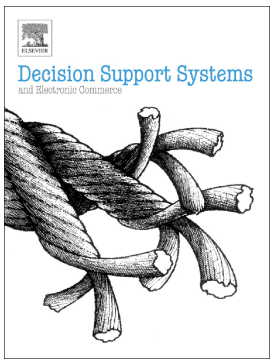

Exploring the use of deep neural networks for sales forecasting in fashion retail

A.L.D. Loureiro, V.L. Miguéis, Lucas F.M. da Silva

PII: S0167-9236(18)30139-8
DOI: doi:10.1016/j.dss.2018.08.010
Reference: DECSUP 12984

To appear in: *Decision Support Systems*

Received date: 31 January 2018
Revised date: 22 August 2018
Accepted date: 22 August 2018

Please cite this article as: A.L.D. Loureiro, V.L. Miguéis, Lucas F.M. da Silva , Exploring the use of deep neural networks for sales forecasting in fashion retail. Decsup (2018), doi:10.1016/j.dss.2018.08.010

Exploring the use of deep neural networks for sales forecasting in fashion retail

By

A.L.D. Loureiro *(corresponding author)*
Faculdade de Engenharia, Universidade do Porto
Address: Rua Dr. Roberto Frias
4200-465 Porto
Portugal
Email: ana.loureiro@fe.up.pt

V. L. Miguéis
Faculdade de Engenharia, Universidade do Porto
Address: Rua Dr. Roberto Frias
4200-465 Porto
Portugal
Email: vera.migueis@fe.up.pt

Lucas F. M. da Silva
Faculdade de Engenharia, Universidade do Porto
Address: Rua Dr. Roberto Frias
4200-465 Porto
Portugal
Email: lucas@fe.up.pt

Exploring the use of deep neural networks for sales forecasting in fashion retail

Abstract

In the increasingly competitive fashion retail industry, companies are constantly adopting strategies focused on adjusting the products characteristics to closely satisfy customers' requirements and preferences. Although the lifecycles of fashion products are very short, the definition of inventory and purchasing strategies can be supported by the large amounts of historical data which are collected and stored in companies' databases. This study explores the use of a deep learning approach to forecast sales in fashion industry, predicting the sales of new individual products in future seasons. This study aims to support a fashion retail company in its purchasing operations and consequently the dataset under analysis is a real dataset provided by this company.

The models were developed considering a wide and diverse set of variables, namely products' physical characteristics and the opinion of domain experts. Furthermore, this study compares the sales predictions obtained with the deep learning approach with those obtained with a set of shallow techniques, i.e. Decision Trees, Random Forest, Support Vector Regression, Artificial Neural Networks and Linear Regression. The model employing deep learning was found to have good performance to predict sales in fashion retail market, however for part of the evaluation metrics considered, it does not perform significantly better than some of the shallow techniques, namely Random Forest.

Keywords: Sales forecasting; Fashion retail; Support vector regression; Artificial neural networks; Deep neural networks

1. Introduction

Fashion retail is a highly competitive market, where inventory control plays a key role in the profitability of companies. Accurate sales forecasting is therefore fundamental to be successful in this environment. If sales forecasting is not accurate, stock-out or overstock situations might occur, which can have a direct and immediate impact on the company's profitability (Agrawal & Schorling, 1996; Sun, Choi, Au, & Yu, 2008; Baecke, De Baets, & Vanderheyden, 2017; Sodero & Rabinovich, 2017; Xia & Wong, 2014) . The effect is not restricted to profitability performance, as quality of the customer service can also be affected by an inefficient forecasting system. For example, if a customer is faced with a stock-out situation, they might decide to shop in a different retailer (Corsten & Gruen, 2004). Additionally, it is known that the fashion industry operates with long supply chains involving a large number of actors (such as raw materials suppliers, manufacturers, distributors and retailers), which leads to orders being placed before there is an accurate understanding of the demand level for the products (H. L. Lee, Padmanabhan, & Whang, 1997; H. Huang & Liu, 2017).

Despite its relevance, sales forecasting is a complex subject because the sales' success of a product is highly dependent on the personal taste of consumers, which varies greatly (Allenby, Jen, & Leone, 1996; Choi, Hui, Liu, Ng, & Yu, 2014). In addition, the lifecycle of fashion products is typically very short, being replaced every new season by new products with no historical sales data (Choi, Hui, Ng, & Yu, 2012). Fashion collections are also composed of an extremely large number of different products in many different sizes, corresponding to many different stock keeping units (SKUs) (Liu, Ren, Choi, Hui, & Ng, 2013). Moreover, several external factors can also have a direct impact on the sales including the weather conditions, holidays, marketing actions, promotions, fashion trends and the current economic environment (Sébastien Thomassey & Fiordaliso, 2006; Sun et al., 2008; Ni & Fan, 2011).

Considering the aforementioned aspects, accurate sales forecasting is therefore crucial to assist in production planning and improving business. These published works also clearly demonstrate the complexity associated with the underlying forecasting process.

This case study intends to contribute to the improvement of sales forecasting problems by:

1. predicting future product sales for new individual SKUs, taking into account a large and highly diverse number of variables which can impact sales performance. In the existent literature, the models proposed to deal with this problem usually include a limited number of factors that influence sales;
2. evaluating the performance of deep neural networks (DNN) to perform predictions in a fashion retail sales forecasting context. The use of deep learning in this context is very incipient and merits further exploration. It is also intended to compare the performance of the DNN model to that of four shallow data mining regression techniques, to better understand if the use of more advanced methods significantly improves the model accuracy or, if simpler methods can achieve similar results;
3. development of predictive models in a fashion retail context, in which the opinion of domain experts and the characteristics of the products are combined and included in the model as part of the predictive variables.

Although this work is focused in the fashion retail industry, the proposed models can also be applied to other retailing domains since the characteristics of the problems are usually similar: a dataset of historical data available to support the forecasts, predictions required for future new products for which no time series data is available, a full registry of the features of the old and new items and no overlap of products between different seasons.

This work is presented in several sections. The following section presents a review of previous works on the use of data analysis techniques for sales forecasting. The case study and the analyzed data are described in the third section. The Methodology section provides a description of the methodology followed and the performance evaluation criteria applied. The fifth section presents all the results obtained as well as a discussion focused on these results. Finally, the conclusions of the work are presented, and possible future works suggested.

2. Related work

In order to explore the issue of fashion retail sales forecasting, several analytical models have been developed by researchers over the years.

Among statistical techniques, those based on time series forecasting methods, e.g. ARIMA, SARIMA, exponential smoothing (Brown, 1962), regression (Papalexopoulos & Hesterberg, 1990), Box & Jenkins (Box & Jenkins, 1976) and Holt Winters (Winters, 1960) are the most commonly employed (Abraham & Ledolter, 2009).

Despite their popularity, statistical methods present several limitations such as the need of converting qualitative data into quantitative, the need to aggregate information or the requirement for time series data, which usually is not available in the fashion industry, where SKUs are replaced very often and sales pattern can be irregular (Hui & Choi, 2016). Consequently, advanced techniques such as fuzzy systems (Zadeh, 1965), or data mining models have emerged to surpass the drawbacks of the statistical methods. These methods are able to handle large datasets and can deal with both numeric and nominal data (Tan, Steinbach, & Kumar, 2006). It is possible to use these methods to model complex, non-linear relationships between sales and the variables of historical items, which can then be generalized to perform sales forecasting of future items (Wong & Guo, 2010), with satisfactory results being reported in diverse studies. As an example, Au, Choi, & Yu (2008) combined neural networks with evolutionary computation to support a forecasting system, showing that the performance of their model is better than that demonstrated by SARIMA model. They report that their model is suitable for short-term retail forecasting of low demand uncertainty and weak seasonal trends. However, a good compromise between accurate results and computing time is difficult to obtain.

In recent years, other techniques have been gaining special attention from the researchers, namely the deep learning techniques. The growing use of deep learning techniques can be attributed to their flexibility and the capability to significantly outperform other data mining methods (LeCun, Bengio, & Hinton, 2015). Deep learning can be superior to simpler data mining techniques (known as shallow methods) for cases with high complexity and large datasets (Larochelle, Bengio, Louradour, & Lamblin, 2009). These techniques have been used to solve real-world problems in diverse research areas such as image classification (Krizhevsky, Sutskever, & Hinton, 2012; Hu, Peng, Yang, Hospedales, & Verbeek, 2018), speech recognition (Dahl, Yu, Deng, & Acero, 2012) and bioinformatics (Lusci, Pollastri, & Baldi, 2013; Xu et al., 2015). The use of deep learning techniques for demand forecasting is also rising, with some studies already published on research topics such as transportation systems (Ke, Zheng, Yang, & Chen, 2017; W. Huang, Song, Hong, & Xie, 2014),

hospital management (Jiang, Chin, Wang, Qu, & Tsui, 2017) and electricity load (Qiu, Ren, Suganthan, & Amaratunga, 2017). However, to the best of our knowledge the application of deep learning algorithms in sales forecasting in retailing is still incipient with a very limited number of published works. An example in retail presents a supermarket sales prediction model, constructed to predict future sales based on the sales of previous days (Kaneko & Yada, 2016). In other work, Lin & Jeffrey (2016) employ a deep learning framework to predict daily number of customers in different points of sales in pharmacy industry.

It becomes evident that the large number of techniques available for the prediction of fashion retail might causes difficulties on the selection of the best prediction technique for a given context. The comparison of the performance of the methods is therefore crucial to understand if the more complex methods offer significantly better performance or, alternatively, if the simpler shallow methods are enough for the task. A few authors have presented comparisons to this purpose applied in other contexts, e.g. Koutsoukas, Monaghan, Li, & Huan (2017) in modelling bioactivity data and Chong, Han, & Park (2017) in stock market analysis and prediction.

Despite the undeniable contribution that the use of data analytics techniques brings to the improvement of forecasts, due to their higher objectivity and higher reasoning capabilities, the inputs coming from the domain experts should deserve special attention. In the literature, there are several studies in different research areas which address the added value brought by the inclusion of the opinion of experts in the development of forecasting models, emphasizing the key role of human input in situations where predictions should be readjusted. These readjustments can occur due to additional information that might become available or the existence of exceptional events (e.g. promotions or news products launches) (Fildes, Goodwin, & Lawrence, 2006; Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009; Baecke et al., 2017). A common finding of these research works is that the performance of forecasting models is better when combining the experts' knowledge with data analytics tools instead of relying on just one of them (Franses & Legerstee, 2011; Coussement, Benoit, & Antioco, 2015). However, in a fashion retail context, no study been found where sales are predicted considering characteristics of the products and the opinion of domain experts as predictive variables.

In what concerns the combination of data mining techniques with human judgment, Sinha & Zhao (2008) investigated the effect of incorporating the human knowledge in seven data mining

classification models in the context of indirect bank lending, having concluded that for all techniques better results are achieved when the domain knowledge is present in the model.

Although there are some studies in the literature in which sales prediction for the fashion industry is made by categories of products (Wong & Guo, 2010; Ren et al., 2015), few studies in which the sales forecasting is performed for individual SKUs can also be found. An exception is, for instance, Sébastien Thomassey & Happiette (2007) which combined clustering and neural networks to develop a decision aid system to predict sales of new items based on their characteristics and similarities with previous items. The authors compare the performance of their model with more basic models including a Naïve Bayesian classifier. Yu et al. in 2011 and Tehrani & Ahrens in 2016 perform similar studies applying different techniques. A common characteristic of these three research works is the consideration of few variables to predict sales (not exceeding a total of 4) with none of the considered variables reflecting the domain knowledge.

This work intends to contribute to the literature by evaluating the efficiency of DNN to perform sales predictions in the fashion retail industry when considering simultaneously aspects of different natures, namely: physical characteristics of the products, their price and features representing the domain knowledge.

3. Case study and data

The company used as case study operates in the fashion market. It has more than 900 stores scattered around the world, but with more incidence in Spain and Portugal. Although this company offers a large variety of products such as umbrellas, bags, watches and hats, the data analyzed in this work is related to the sales of 684 types of women bags, during the seasons Spring-Summer 2015 (SS15) and Spring-Summer 2016 (SS16).

The data made available for this study is comprised of historical sales data of each product and its physical characteristics, logistical and internal organizational aspects of the company and even the opinion of domain experts, which is believed to have great influence on the sales reached by the products. All these factors were used to understand which characteristics make a product more

attractive for the clients and thereby assist the purchasing department to more precisely define the quantity of product to be ordered.

Currently, the quantity of bags ordered by the company is determined without any kind of quantitative model as support, by taking into account three factors: sales of similar products in previous and homologous seasons, experience of the purchasing department staff, development staff and marketing department and information related to future trends collected from different sources, such as fashion blogs and fashion shows. Table 1 shows the type of information collected, that was the basis for this study. For each variable, the name, the type of variable it represents, its meaning, and the total number of categories it can assume are specified. This table also includes a classification of the variables, considering their nature. They were classified in the following classes: Product characteristics (PC), Logistical and internal organizational aspects of company (LIO) and Domain experts (DE). A set of 10 variables characterize the products, with the price of the product being the only numerical variable. All other variables are categorical.

Table 1 - Variables currently collected by the company.

Name	Type	Description	Variables classification	Number of categories
Family	Categorical	Identifies the type of material and/or pattern used in product production	PC	18
Subfamily	Categorical	Identifies the format of the product	PC	11
Color type	Categorical	Identifies if it is a single-color or multi-color product	PC	2
Color	Categorical	Color of the product	PC	30
Fashion	Categorical	Related to season's trends and to the number of weeks that the product is available in store for sales. The products are classified according to its proximity to the current season's trends	DE / LIO	4
Segment	Categorical	Target group	DE	2
Store type	Categorical	Types of store where the product will be available for sales.	DE / LIO	4
Price	Numerical	Product price	PC	-
Size	Categorical	Product size	PC	4
Expectation level	Categorical	Sales expectation level	DE	4
Sales	Numerical	Total number of units sold		-

To complement the short description presented in Table 1, it should be clarified that the company classifies its stores into four different types (A, B, C or D), a rating that is related to the size of the

store and its sales potential. A store classified as ‘A’ is a store with a large display area and a high sales potential which will receive the totality of the products developed for the season. On the other hand, a more limited variety of products will be available in smaller stores with lower sales potential, classified as ‘D’. This classification of the stores is used to define the total amount to be ordered from each item. To infer this quantity, agglomerations of stores typologies are created, based on the following methodology: the stores classified as ‘D’ will only be supplied with products bought to be displayed in all stores of the company, while ‘A’ stores can receive more specific products, as these stores have demonstrated the ability to sell them. As a result of this process, a product classified as DCBA in the variable ‘Store type’ is one that will be available for sales in all the company stores, while a product whose ‘Store type’ is A is only displayed in stores classified as ‘A’. For the ‘Expectation level’ predictor, there are four possible values: ‘SB’ corresponds to the highest value of sales expectations, ‘M1’ represents normal sales expectations, ‘M2’ is used for products with average sales expectations while the last classification, ‘M3’, is reserved for products with low sales expectations. This classification is defined by the marketing, development and purchasing departments and is related to the sales expectation level of the product. Additional explanation is also required for ‘Fashion’ predictor as the category names do not convey the differences between each category. This variable groups the products taking into account fashion and season’s trends. The articles are classified according to their proximity to the current season’s trends from ‘Trendy’ to ‘Basic’, with ‘Trendy’ products being designed to be closer to the current season’s trends. The other possible classification of this attribute is ‘Distribution Centralized’, which is similar to ‘Basic’, differing only in the distribution strategy employed. This classification is also related to the number of weeks that the product is available in store for sales. Products classified as ‘Trendy’ are available in stores for 4 weeks and ‘Basic Fashion’ for 6 weeks. A longer display period is reserved for the ‘Basic’ and ‘Distribution Centralized’ categories, which are available for 8 weeks. The products are also distinguished according to the target audience for which they are designed (variable ‘Segment’). Products classified as ‘Teen’ are designed for teenagers, while products intended to be sold to older women are classified as ‘Woman’. This classification represents the opinion the marketing department’ staff in what concerns the age of people who will demonstrate a greater interest in the product. However, this

classification does not restrict the sale of a product to a target group and all products can be purchased by anyone. In Appendix A all the possible categories for each independent variable are presented.

As previously mentioned, the 'Expectation level', 'Store type', 'Fashion' and 'Segment' independent variables classify the products according to the sales level that the company's staff believes the product can reach, to the aggregation of stores in which the product should be displayed, to the product's proximity to the season's trends and to the products' target group, respectively. All these four classifications are generated by the company' specialized staff, which in this case study represent the domain knowledge. In line with what was found in the literature, with several studies proving the added value of integrating the domain knowledge into forecasting models, these three variables will be considered for the construction of the sales prediction models.

4. Methodology

In this section a description of the methodology followed is presented and the performance evaluation metrics applied to evaluate the performance of the models are identified. A brief description of the data analytic techniques employed is also given.

4.1 General description

This research work has the objective of predicting the sales of new fashion products using data mining techniques, while also aiming to explore the potential of a deep learning approach in this context.

Although there is no historical sales data for products newly introduced in each season, companies possess information regarding the sales of previous products and have complete knowledge of the characteristics of historical and new items. The prediction of sales of future items can therefore be achieved based on the sales of historical items and the characteristics of old and new items, exploring their similarities. For this purpose, this work proposes to explore different regression data mining techniques to develop a model to forecast the sales of new products.

In this model, the variables which characterize the products, those related to logistical and internal organizational aspects of the company and those that reflect the decisions of domain experts are employed as independent variables to predict the sales of a product (dependent variable) in a given

season. For the learning process of the model, historical sales data of the products corresponding to a homologous season is used.

The methodology followed in this study is synthesized in Figure 1 and is described in detail in the following paragraphs.

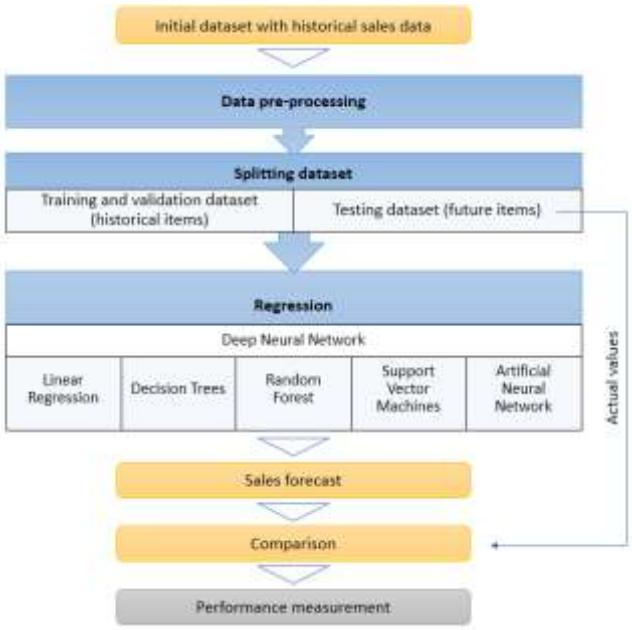

Figure 1- Schematic representation of the methodology followed.

As illustrated in Figure 1, after a data pre-processing step, the initial dataset was split into two sub datasets according to the season to which the observation belongs to. The information regarding the SS15 season is the basis for the training procedure, which corresponds to a dataset composed of 384 items for which historical data is available. The data corresponding to the SS16 season is used to perform the performance evaluation of the models, i.e. a total of 300 future items. It is important to highlight that the products do not transit from one season to the following homologous season, which means that there is no overlap of products between two different seasons.

A summary of the volume of sales for the SKUs belonging to the SS15 and SS16 seasons is graphically represented in Figure 2 and Figure 3.

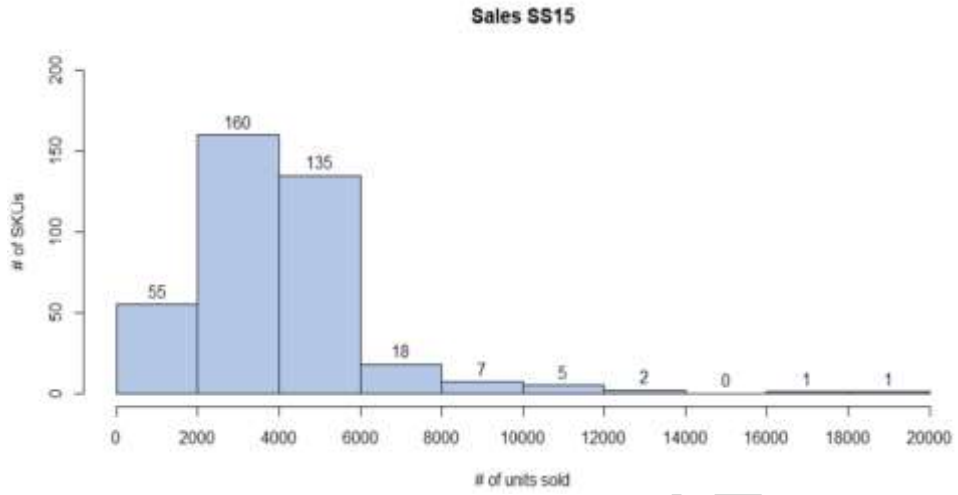

Figure 2 - Sales of season SS15.

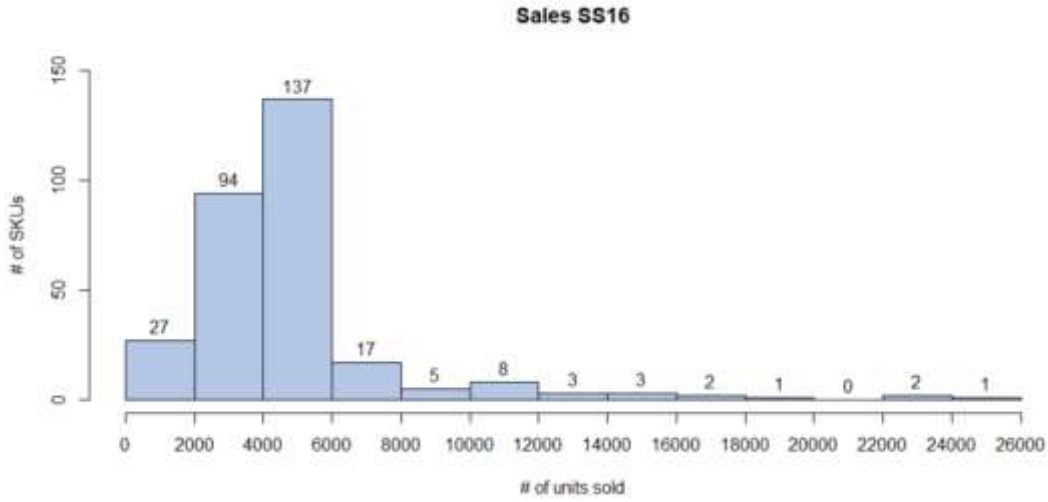

Figure 3 - Sales of season SS16.

Figure 2 and Figure 3 show that in the two seasons considered a large portion of the products recorded sales ranging from 2000 to 6000 units. For both seasons this portion corresponds to about 77% of the total number of SKUs which constitute the respective collection. It is also possible to conclude that a few SKUs can be considered as outstanding products as their sales outperform the sales of remaining products, reaching more than 14000 sold units.

A more detailed characterization of the datasets per each predictive category is presented in Appendix A, including the total number of products, the total number of units sold and the mean value of sales.

The construction of the predictive models encompassed three important steps which were connected to each other, improving the models' learning process and leading to the creation of more precise forecasting tools. These steps are: application of a greedy approach to select the predictive variables

integrated in the model, an optimization process of the parameters involved in each technique and a testing process of the model. A description of each of these steps and the way how they are chained during the model's development is presented in the next paragraphs. This description is supported by the presentation of the algorithm' structure in Figure 4.

```
Partitioning of the training dataset into ten complementary subsets (folds) applying a cross-validation
WHILE the current R2 is higher than the last R2 DO
    FOR each combination of variables (current set of predictors and a new variable)
        FOR each possible value of parameter_1
            FOR each possible value of parameter_n
                FOR each fold
                    Train the predictive model on nine folds
                    Compute predictions for the remaining fold (validation set)
                    Calculate the R2 between predictions and the values of the validation set
                END FOR
                Get the mean value of R2 of the ten folds for the pair (combination of variables / combination
of parameters)
            END FOR
        END FOR
    END FOR
    Store the highest value of R2 as the current R2
END WHILE

Store the combination of predictive variables and parameters that result on the best performing model

FOR ten iterations
    Create a sampling with reposition of the initial training set
    Create a sampling with reposition of the initial testing set
    Train the predictive model on the sample training set
    Compute predictions for the sample testing set
    Calculate the R2 between predictions and the values of the testing set
    Calculate the remaining performance metrics of the model
END FOR

Determine the final performance metrics of the model as the mean value of the ten iterations
```

Figure 4 - Algorithm's structure.

The aim of this procedure is to firstly determine the selection of the variables and parameters which lead to better performing models using only historical data and then, evaluate the performance of the best model achieved when tested on future and unseen information.

For this purpose, the initial training dataset was partitioned into K folds by the application of a *k-fold cross-validation* with K equal to ten. The best parameters and predictors combination were found by repeating the training-validation procedure multiple times (as many as the number of defined folds) with each of these folds being left out at a time for training-validation sequence.

For the selection of the best combination of predictive variables to be considered on the model, a greedy feature selection approach was applied making a locally optimal choice at each stage. To start the process, the dependent variable was associated to each of the independent variables separately, i.e., the process begins considering all the possible combinations obtained with just one single independent

variable (in a total of 10 possibilities). For each of these combinations a parameters optimization process was performed which incorporated a model testing process on the validation dataset. The optimization of parameters consisted of a brute-force grid search of the defined values for the parameters under analysis. The value of the coefficient of determination (R^2) for each pair (combination of variables / combination of parameters) was determined as the mean value of the ten R^2 values obtained during the model's evaluation process. The best combination of parameters for each combination of variables is chosen as the one that maximizes the R^2 mean value. Analogously, the best combination of variables is selected as that which allows to reach the highest value of mean R^2 among all the possible variable combinations.

The individual predictor that enabled to achieve the highest mean R^2 is then combined with the remaining ones. The process is repeated until there is no improvement in the mean R^2 value, which means that the inclusion of an additional predictor (regardless of which predictor it is) does not improve the R^2 between the predictions estimated with the model and the observed sales values. The combination of predictors and set of parameters that resulted in the best performing model is then used on the construction of the model to be used to predict future sales.

The performance evaluation process of the final model follows a bootstrapping approach (Efron, 1979) and consisted on the repetition of the following steps ten times: creation of a training and testing datasets (with the same size of the original datasets) resultant from sampling with reposition of the original SS15 and SS16 datasets respectively, development of the predictive model, determination of the predictions for the testing dataset and calculation of the R^2 and the remaining performance measures between the predictions and the testing dataset under consideration.

With the adoption of this repetitive sampling procedure in the construction and testing of the predictive models, the models' robustness increases and allows to avoid overfitting.

Five different widely used evaluation metrics were calculated to measure the performance of the final predictive models. These metrics are R^2, root mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE) and mean squared error (MSE).

The final values for the performance metrics of the models were determined as the mean value of the ten iterations.

4.2 Data mining techniques

As mentioned before, the objective of this study is to explore the use of DNN (Smolensky, 1986) to predict product sales. Additionally, and to allow a comparison of the performance of deep learning techniques with shallow techniques, a linear regression and four different shallow data mining techniques were also employed: Decision Trees (Quinlan, 1986), Random Forest (Breiman, 2001), Support Vector Regression (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998) and Artificial Neural Networks (McCulloch & Pitts, 1943).

As previously mentioned, some of the data mining techniques employed require the definition of a set of parameters that can be tuned, in order to determine the value for each parameter that leads to improvements of the efficiency of the forecasting model. The tuning of the parameters for all the employed techniques was achieved by the application of a brute-force grid search (or exhaustive search). Linear regression was the only technique for which no parameter optimization was performed.

A brief description of the applied techniques, together with the parameters chosen for optimization and the correspondent explored values are presented below for each of the employed techniques, starting with the shallow techniques:

Decision Trees

Decision trees (DT) is a technique commonly used to build regression or classification models in the form of a tree structure. The high popularity of this technique is due to its ease of use and interpretability by users. Besides this, nominal variables can be treated, and the technique is independent from the variable scale. However, the performance of these models is usually below the standard levels and they are considered as not very robust (Tan et al., 2006).

For this technique, the parameters chosen to be tuned were the minimum number of observations in any terminal node of the tree and the complexity parameter. For the first parameter, values from 2 to 20, with a step of 1 were tested while for the complexity parameter values of 0 and 0.01 were considered.

Random Forest

This technique derives from the Decision Trees (DT) and was developed to overcome the weaknesses revealed by DT allowing to improve the accuracy of DT and bypassing problems such as the high sensitivity to small variations in data. It is based on an ensemble of trees followed by the calculation of the mean value of the predictions obtained at the final node of each tree and avoids the lack of robustness demonstrated by a single decision tree. In this approach, each tree is grown by using a subset of independent variables randomly selected. As is the case for DT, Random Forest (RF) is considered a simple method (Breiman, 2001).

Concerning the parameters tuning, the number of trees to grow was fixed at 1600 while for the number of variables randomly sampled at each split, values ranging from 1 to the number of predictors considered in the variables' combination under analysis were tested.

Support Vector Regression

Support vector regression (SVR) is a category of support vector machines that aims at mapping the original data into a high dimensional feature space. In that space, the support vector regression algorithm finds the best regression hyperplanes, allowing to estimate the value of the dependent variable (Witten et al., 2011).

The influence of two different *kernel* types (linear and radial) and the different parameters (cost and gamma) required for each kernel were also studied. For the parameter cost, values of 2^n with n between 2 and 9, with a step of 1 were considered, while for gamma the boundary tested values were 0.5, 1 and 2.

Artificial Neural Networks

ANN are a set of tools constructed by simple processing units (neurons) connected with each other, where each connection has a weight associated with it. In these networks, the neurons are organized in layers where each neuron receives a set of inputs (corresponding to the outputs of the neurons from previous layer) and outputs a nonlinear weighted sum of its inputs to the neurons in the next layer to which it is connected. These weighted connections allow the model to "learn" the relationships between variables. A single feed-forward neural network is composed of an input layer, a hidden layer

and an output layer. For neurons in hidden and output layers, an activation function is applied to its input values (Han, Pei, & Kamber, 2011).

Many kinds of neural networks and neural network algorithms have been developed. Backpropagation is the most popular neural network algorithm and it is the one applied in this study.

The parameters tuning of ANN involved several parameters. Regarding the configuration of the hidden layer, the number of hidden neurons ranged between 50 and 200 with a step of 10. For the nonlinear activation function, two different possibilities were tested, a rectifier linear function and a tanh activation function which is a rescaled and shifted logistic function. Concerning the regularization, two different methods were used, and their coefficients were set to 1e-5. One of these methods constrains the absolute value of the weights while the other constrains the sum of the squared weights. To speed up the convergence of the model, an adaptative learning rate algorithm called ADADELTA (Zeiler, 2012) was adopted. This algorithm dynamically adapts over time and requires no manual tuning of a learning rate. However, two parameters need to be tuned. The first one relates to the memory to prior weight updates and was tuned for the following values: 0.9, 0.95, 0.99 and 0.999. The second parameter of ADADELTA is identical to learning rate annealing during the initial training and can be considered as analogous to momentum at later stages, enabling forward progress. The tested values for this parameter were: 1e-10, 1e-8, 1e-6 and 1e-4. Parameters related to early stopping of the model were also defined, based on the value of the mean squared error. It was established that the model building should stop if the MSE on the validation set does not improve (decreases) by more than 1% at all for 5 consecutive scoring epochs. It was also defined that the dataset should be iterated ten times.

Deep Neural Networks

DNN are more complex, fully connected ANN composed by more than one hidden layer where each successive layer uses the output from the previous layer as input (Pal & Mitra, 1992). Although different deep architectures are available, the most common are the feed-forward networks with application of the backpropagation algorithm for learning process together with an optimization technique (Gardner & Dorling, 1998; Koutsoukas et al., 2017). DNN are structured so that each layer is trained on a limited set of features, resulting from the previous layer's output. While the initial

layers can only process simple features, later layers are able to recognize more complex data, achieving this by combining features resulting from the processing performed by the previous layers. Although they exhibit good performance, DNN are prone to overfitting, which can be decreased by applying a regularization technique such as weight penalty, early stopping or dropout during training (Bengio, 2009; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014; LeCun, Bengio, & Hinton, 2015).

In what concerns the deep learning, a feedforward neural network architecture was adopted, and a stochastic gradient descent applied as training algorithm, in which the gradients are computed via backpropagation. The number of hidden layers was defined as three, with all three layers containing the same number of neurons. Four different configurations were tested, using 50, 80, 100 and 200 neurons per layer. In order to align the tuning of the parameters of ANN and DNN as much as possible, for DNN the values tested for activation function, for parameters related to regularization, adaptive learning rate and early stopping of the model were set to be same as those used for ANN, and which were already introduced previously.

The forecasting performance of the techniques applied in this study was evaluated considering the previously described method, i.e. the mean value of R^2 obtained from the ten iterations of the testing dataset (resulting from a sampling of the same size with reposition of the dataset containing the sales of SS16 season) to estimate the quality of the classifier. For this purpose, real sales data was compared to the predicted values resultant from the application of the proposed model.

4.3 Information fusion-based sensitivity analysis

Besides the comparison of the performance of different predictive models, in this work the relative importance of each of the independent variables is also evaluated. For this purpose and although there are several methods to infer the importance of variables in the literature (Cortez & Embrechts, 2013), in particular, in this study it was decided to perform a sensitivity analysis by the application of the well-established information fusion based sensitivity analysis technique. Using this technique the importance of the variables is computed considering the contribution of all analytical models, allowing the reduction of the bias and uncertainty associated to individual prediction models (Oztekin et al.,

2013). In this technique, the contribution of each independent variable on the dependent variable is measured through the evaluation of the change in the performance of the model caused by the absence of that predictive variable. Therefore, the measure of sensitivity of a particular predictor can be assessed by the ratio between the performance of the model without this predictor and the performance of the model including this predictor (Delen, Oztekin, & Tomak, 2012). Consequently, the greater the decrease in the performance of a model due to the absence of a predictor, the greater the importance of this predictor because the higher sensitivity of the model to this specific predictor. The sensitivity measure expressed by Eq (2) has been demonstrated to be an appropriate measure to rank the predictors according to their importance value (Saltelli, Tarantola, Campolongo, & Ratto, 2004) and is commonly applied. In this expression F_t refers to the output variable and $V(F_t)$ corresponds to the unconditional output variance, the expectation operator E represents the integral over all input variables excepting the variable X_i while the variance operator V implies a further integral over X_i.

$$S_i = \frac{V_i}{V(F_t)} = \frac{V\left(E(F_t|X_i)\right)}{V(F_t)} \quad (1)$$

$$S_{n(fused)} = \sum_{i=1}^{m} \omega_i S_{in} = \omega_1 S_{1n} + \omega_2 S_{2n} + \ldots + \omega_m S_{mn} \quad (2)$$

In accordance to the fusion based sensitivity analysis technique, an information-fusion based sensitivity measure of the variable *n* with *m* prediction models can be obtained by Eq (2).

In the previous equation, ω_i represent the normalized weighting coefficient of each individual prediction model (meaning that $\sum_{i=1}^{m} \omega_i = 1$) and S_{in} is the sensitivity measure of the *nth* predictor in the *ith* model. Regarding the weighting coefficients, the better the predictive performance of a prediction model, the higher their respective weight in the fusion function, which means that well performed models will have a higher contribution than the models which does not have such good performance (Sevim, Oztekin, Bali, Gumus, & Guresen, 2014).

5. Results and discussion

Technical results

In this section the results obtained during the study are presented and discussed. Table 2 summarizes the best results achieved for each of the regression techniques applied according to the different evaluation metrics employed.

Table 2 - Regression techniques performance on testing dataset

		Regression technique					
		DNN	**DT**	**RF**	**SVR**	**ANN**	**LR**
Evaluation metric	**R^2**	0.727	0.687	**0.756**	0.672	0.697	0.568
	RMSE	**1861.0**	1966.5	1921.8	2030.6	1909.4	2890.5
	MAPE	0.378	0.377	**0.345**	0.393	0.387	0.451
	MAE	1141.4	1152.1	**1102.8**	1255.2	1110.1	1621.9
	MSE	**3526377**	3921105	3746350	4177228	3715802	8460688

The same performance metrics where also determined when the model was applied to the validation dataset. The results are reported in Appendix B.

The results show that, with the exception of the linear regression, for all applied regression techniques, the proposed models have a high capability to predict the sales of a product ($R^2 \geq 67\%$). For that reason, the proposed models (even when supported by the shallow techniques) can be considered as powerful tools for supporting the decision-making process of the company. The data of Table 2 also allows to conclude that none of the considered techniques exhibits higher performance in all the metrics explored and therefore no single technique can be considered as the best. The definition of which technique performs better depends on which evaluation metric the decision is based on. If R^2 is used as the accuracy measure, RF can be considered as the best performing technique, otherwise if the decision is based on error related metrics, RF shares the leadership with DNN as each of these techniques presents the best results for two out of four metrics.

These results demonstrate that in spite of being recommended for analysing databases containing large amounts of data, DNN can also perform well when applied to smaller datasets, outperforming most of the shallow techniques. The results also show that the application of shallow techniques enables to

obtain accurate forecasts. Indeed, high values of R^2 are obtained in almost all the applied shallow techniques (excepting LR).

The best performing neural networks are typically those structured with multiple hidden layers, where the learning process is confined to a subset of features per layer, enabling an increase of the learning ability of the network and consequently the accuracy of feature extraction. However, the results of the DNN and ANN techniques show just a slight improvement with the inclusion of additional hidden layers.

From a managerial point of view, the choice of the best technique should represent a balance between the performance of the models, their interpretability and their comprehensibility. In this particular case, the adoption by the company of a decision support tool based on DNN may involve a significant effort to understand the technique and the parameters tuning involved. In opposition, techniques such as RF represents a very comprehensible technique, while providing relatively satisfactory results and insights on the impact of the predictive variables on the performance of the models.

As previously described, the construction of the forecasting models involved a greedy approach where, for each step, the inclusion of an additional predictive variable only occurs in cases where this inclusion represents an added value for the model's performance. Consequently, for each technique the combination of variables which exhibit better results are different, which is due to specificities of each algorithm. The combination of variables resultant from each technique is shown in Table 3. In accordance to the literature, for this greedy approach all types of predictive variables were made available for the construction of the models, namely the independent variables related to physical characteristics of the products and those linked to the domain knowledge.

Table 3 - Best performing combination of variables for each technique.

Technique	**Combination of variables**
DNN	Expectation level + store type + subfamily + size + family + color
RF	Expectation level + store type + family + segment + color type
SVR	Expectation level + store type + price + subfamily + family + color
ANN	Expectation level + store type + subfamily + family + color + segment + price
DT	Expectation level + store type + family + segment + color type
LR	Expectation level + store type + fashion

The results obtained with the feature selection approach adopted show that 'Expectation level' and 'Store type' are the two predictive variables included in all of the six data analytical techniques

employed, which demonstrates that the knowledge of the domain experts plays a key role on sales prediction. The 'Expectation level' variable captures the sensitivity of the experts on the preferences of the customers and consequently was expected to have a high impact on sales prediction. In line with this, the 'Store type' assigned to each product was also expected to be a relevant variable. Products allocated to all types of stores are expected to have higher sales than those allocated only to one type of store.

'Family', 'Subfamily', 'Color' and 'Segment' are the other variables which appear the most on the best performing combinations of variables, according to the greedy procedure used. 'Family', 'Subfamily' and 'Color' naturally reveals that the physical characteristics of the product impact its sales potential, while 'Segment' reinforces the importance of role of the domain experts' opinion for achieving good results.

As previously mentioned, this work also allows an insight on the significance of the predictors on product sales with the information fused-based sensitivity analysis being the technique employed to perform such analysis. The results are graphically represented in Figure 5, where the predictors are listed in descending order of importance and the bar correspondent to each predictor is colored in agreement to the group of variables it belongs to (product's physical characteristics, logistical internal organizational aspects of the company or domain knowledge).

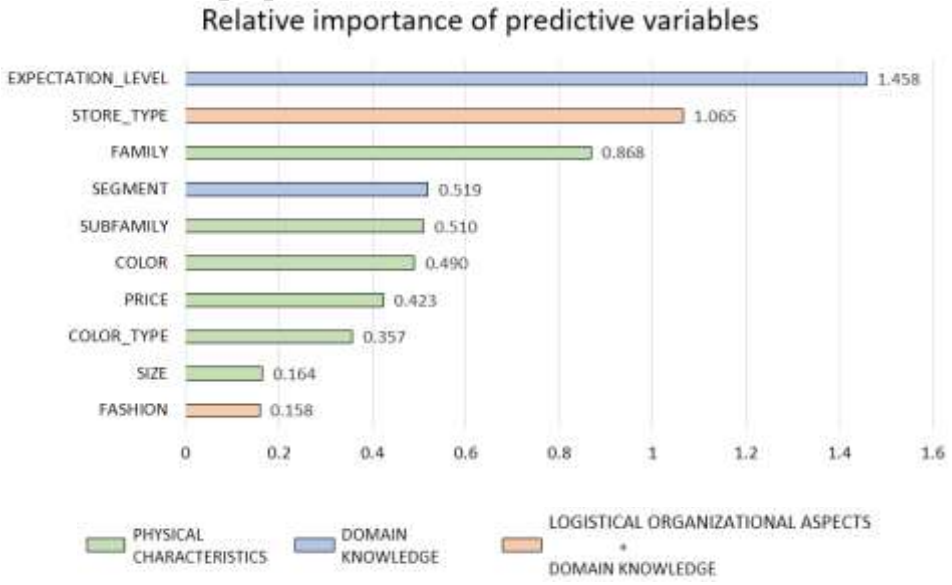

Figure 5 - Information fused sensitivity results.

The results presented in Figure 5 show 'Expectation_level' as the most important variable in determining the number of sales of a product, which means that, if only this variable is excluded from the model, this will have a more negative effect on the predictive capability of the models than the exclusion of other variables, individually. To accurately assess the sensitivity of the predictive models to 'Expectation_level' variable, the performance of all models excluding this variable was determined and the results are presented in the Appendix C. These results emphasize the relevance of this variable, as the performance of the models decreases significantly, which is reflected by the decrease of R^2 and the increase of all error related metrics. This trend is found for all techniques. Thus, once again, these results make clear the importance of the sensitivity of the marketing, development and purchasing departments in determining the potential of sales of each product.

Furthermore, from the analysis of Figure 5, the big impact of all predictors related to the domain knowledge on the models' accuracy is also noticeable, which is reflected on the fact that among the four most impactable predictors, three of them belong to this group of variables. In what concerns the last six predictors to which the predictive models are more sensitive, five of them are related to physical characteristics of the products and the other one, and least important variable ('Fashion'), combines the domain knowledge and logistical and internal organizational aspects of company.

Managerial implications

Overall, the results demonstrate that these models can serve as valid tools to predict sales and help define ordering and marketing strategies in the fashion industry, but managers must be fully aware of a few particularities associated with the use of these models.

Firstly, managers need to understand the complexities associated with the different types of models available, and the different resources necessary. For example, this work demonstrates that DNN can operate successfully with both large and small datasets, but in practice this technique can be harder to implement due to the need for a complex training process. For many cases, simpler techniques such as RF can also achieve good results which might represent a most cost-effective approach.

Another key aspect with practical implications evidenced by this work is the extremely important role played by the experts in the definition of the potential sales. It is undeniable that their sensitivity to the market response is something of extreme value and cannot be fully replaced by the use of models

which disregard this type of information. Managers must ensure that their staff is fully aware of the market and its needs and should integrate their input and knowledge in the decision process. It is also relevant to note that one recognized characteristic of the fashion retail market is its high dependency on the customers' preferences, which are very volatile. For that reason, managers who decide to use this type of models to assist in their decision-making processes must be aware of the necessity of retraining the models every year with new information, as this step is crucial in order to enable the model to capture the most recent preferences of the customers.

6. Conclusions and Future work

Sales forecasting in the fashion retail industry is a complex problem which requires innovative and efficient tools, especially in cases in which the predictions are performed for new products characterized by the absence of sales historical data. In an increasingly competitive industry dependent on diverse and, in certain cases uncontrollable factors, assertiveness in predicting the sales of products is a crucial factor to improve the results of a company. More accurate forecasting tools can support companies on determining, in a more precisely manner, the quantity of each product that should be ordered from the suppliers, contributing positively to the enhancement of the purchasing process. This process will ultimately be reflected in an increase of the profitability, as expensive storage costs of unsold stock are avoided, and the company can operate in a leaner, more efficient manner. A close control of inventory levels in physical stores can also act as a differentiator factor between competitors, as costumers will favor stores which carry in stock the products they desire to acquire, which results in increase customer satisfaction levels and loyalty.

This work undertook the development of a forecasting model to estimate the sales of future products for which no historical data exists. The potential of using a deep learning approach to accomplish this goal was explored. The sales data provided by a company operating in the fashion retail market was used. The dataset includes information on the sales of products from previous collections and the predictive variables characterizing the previous and future products.

The results demonstrate that the use of DNN and other data mining techniques for performing sales forecasting in the fashion retail industry when there is no historical sales data is very promising. The

models proposed may constitute an important tool to assist managers in their products acquisition process. In particular, DNN outperformed the remaining techniques in part of the performance techniques considered. Therefore, although usually suggested as appropriate for the analysis of large databases, DNN techniques can also perform well when applied to smaller datasets. However, it is important to stress that in practice this technique might not be the most suitable, as its training process is more complex when compared with other simpler techniques that still exhibit satisfactory predictive capability, such as RF.

This work has also demonstrated that this type of models and techniques are able to provide valuable insights into the relationships between product attributes and sales using the data which in many cases is already stored and can be easily accessed by the companies. In particular, the results attest to the significance of the variables representative of the domain knowledge, demonstrating the importance of informed human judgement in the decision process. The consumers' preferences are not constant and even the most recent information (closest homologous season) can not exactly reflect the current trends, reinforcing the idea that domain knowledge can guide the purchasing department to predict with high accuracy the demand and adjust orders to suppliers accordingly.

Regarding future work, the authors believe that the results achieved can be improved if other predictive variables are integrated. As an example, the influence of external factors such as the emergence in the market of other competitors can also be studied. The information contained in other sources such as fashion blogs can also contribute to capture the future fashion tendencies. For this purpose, text mining techniques can be employed, enhancing the work currently performed by the company's managers.

The authors also consider that a performance evaluation of the same techniques, on the same context but in the presence of historical data, would be a valuable task. In fact, it seems relevant to analyze the potential of these models in situations in which time series data is available, as in the case of continuity products.

References

Abraham, B., & Ledolter, J. (2009). *Statistical methods for forecasting* (Vol. 234). John Wiley & Sons.

Agrawal, D., & Schorling, C. (1996). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing*, *72*(4), 383–407. https://doi.org/10.1016/S0022-4359(96)90020-2

Allenby, G. M., Jen, L., & Leone, R. P. (1996). Economic trends and being trendy: The influence of consumer confidence on retail fashion sales. *Journal of Business & Economic Statistics*, *14*(1), 103–111.

Au, K.-F., Choi, T.-M., & Yu, Y. (2008). Fashion retail forecasting by evolutionary neural networks. *International Journal of Production Economics*, *114*(2), 615–630. https://doi.org/10.1016/j.ijpe.2007.06.013

Baecke, P., De Baets, S., & Vanderheyden, K. (2017). Investigating the added value of integrating human judgement into statistical demand forecasting systems. *International Journal of Production Economics*, *191*, 85–96.

Beheshti-Kashi, S., Karimi, H. R., Thoben, K.-D., Lütjen, M., & Teucke, M. (2015). A survey on retail sales forecasting and prediction in fashion markets. *Systems Science & Control Engineering*, *3*(1), 154–161. https://doi.org/10.1080/21642583.2014.999389

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, *2*(1), 1–127.

Bogaert, M., Ballings, M., & den Poel, D. V. (2018). Evaluating the Importance of Different Communication Types in Romantic Tie Prediction on Social Media. *Annals of Operations Research*, *263*(1–2), 501–527.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden-Day.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, R. G. (1962). *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs: Prentice-Hall.

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

Choi, T.-M., Hui, C.-L., Ng, S.-F., & Yu, Y. (2012). Color trend forecasting of fashionable products with very few historical data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 1003–1010.

Chong, E., Han, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, *83*, 187–205.

Corsten, D., & Gruen, T. W. (2004). Stock-Outs cause Walkouts. *Harvard Business Review*, *82*(5), 26–28.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, *225*, 1–17.

Coussement, K., Benoit, D. F., & Antioco, M. (2015). A Bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection. *Decision Support Systems*, *79*, 24–32.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio Speech and Language Processing*, *20*(1), 30–42. https://doi.org/10.1109/TASL.2011.2134090

Delen, D., Oztekin, A., & Tomak, L. (2012). An analytic approach to better understanding and management of coronary surgeries. *Decision Support Systems*, *52*(3), 698–705.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, *42*(1), 351–361.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23.

Franses, P. H., & Legerstee, R. (2011). Combining SKU-level sales forecasts from models and experts. *Expert Systems with Applications*, *38*(3), 2365–2370.

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (Third Edition). Elsevier.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28.

Hu, G., Peng, X., Yang, Y., Hospedales, T. M., & Verbeek, J. (2018). Frankenstein: Learning Deep Face Representations Using Small Data. *IEEE Transactions on Image Processing*, *27*(1), 293–303. https://doi.org/10.1109/TIP.2017.2756450

Huang, H., & Liu, Q. (2017). Intelligent Retail Forecasting System for New Clothing Products Considering Stock-out. *Fibres & Textiles in Eastern Europe*, *25*(1), 10–16. https://doi.org/10.5604/12303666.1227876

Huang, W., Song, G., Hong, H., & Xie, K. (2014). Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, *15*(5), 2191–2201.

Hui, P. C. L., & Choi, T.-M. (2016). Using artificial neural networks to improve decision making in apparel supply chain systems. In *Information Systems for the Fashion and Apparel Industry* (pp. 97–107). Elsevier.

Jiang, S., Chin, K.-S., Wang, L., Qu, G., & Tsui, K. L. (2017). Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. *Expert Systems with Applications*, *82*, 216–230. https://doi.org/10.1016/j.eswa.2017.04.017

Kaneko, Y., & Yada, K. (2016). A Deep Learning Approach for the Prediction of Retail Store Sales. In C. Domeniconi, F. Gullo, F. Bonchi, J. DomingoFerrer, R. BaezaYates, Z. H. Zhou, & X. Wu (Eds.), *2016 IEEE 16th International Conference on Data Mining Workshops (icdmw)* (pp. 531–537). New York: Ieee.

Ke, J., Zheng, H., Yang, H., & Chen, X. (2017). Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C-Emerging Technologies*, *85*, 591–608. https://doi.org/10.1016/j.trc.2017.10.016

Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, *9*, 42. https://doi.org/10.1186/s13321-017-0226-y

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, *10*(Jan), 1–40.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*(4), 546–558.

Lin, K.-Y., & Jeffrey, J. P. T. (2016). *A Deep Learning-Based Customer Forecasting Tool*. New York: Ieee.

Liu, N., Ren, S., Choi, T.-M., Hui, C.-L., & Ng, S.-F. (2013). Sales Forecasting for Fashion Retailing Service Industry: A Review. *Mathematical Problems in Engineering*, 738675. https://doi.org/10.1155/2013/738675

Lusci, A., Pollastri, G., & Baldi, P. (2013). Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of Chemical Information and Modeling*, *53*(7), 1563–1575.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Ni, Y., & Fan, F. (2011). A two-stage dynamic sales forecasting model for the fashion retail. *Expert Systems with Applications*, *38*(3), 1529–1536.

Oztekin, A., Delen, D., Turkyilmaz, A., & Zaim, S. (2013). A machine learning-based usability evaluation method for eLearning systems. *Decision Support Systems*, *56*, 63–73.

Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on Neural Networks*, *3*(5), 683–697.

Papalexopoulos, A. D., & Hesterberg, T. C. (1990). A regression-based approach to short-term system load forecasting. *IEEE Transactions on Power Systems*, *5*(4), 1535–1547.

Qiu, X., Ren, Y., Suganthan, P. N., & Amaratunga, G. A. J. (2017). Empirical Mode Decomposition based ensemble deep learning for load demand time series forecasting. *Applied Soft Computing*, *54*, 246–255. https://doi.org/10.1016/j.asoc.2017.01.015

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Ren, S., Choi, T.-M., & Liu, N. (2015). Fashion Sales Forecasting With a Panel Data-Based Particle-Filter Model. *Ieee Transactions on Systems Man Cybernetics-Systems*, *45*(3), 411–421. https://doi.org/10.1109/TSMC.2014.2342194

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity analysis in practice: a guide to assessing scientific models*. John Wiley & Sons.

Sevim, C., Oztekin, A., Bali, O., Gumus, S., & Guresen, E. (2014). Developing an early warning system to predict currency crises. *European Journal of Operational Research*, *237*(3), 1095–1104.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems*, *46*(1), 287–299.

Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory*. COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE.

Sodero, A. C., & Rabinovich, E. (2017). Demand and Revenue Management of Deteriorating Inventory on the Internet: An Empirical Study of Flash Sales Markets. *Journal of Business Logistics*.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Sun, Z.-L., Choi, T.-M., Au, K.-F., & Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, *46*(1), 411–419. https://doi.org/10.1016/j.dss.2008.07.009

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Addison Wesley.

Tehrani, A. F., & Ahrens, D. (2016). Enhanced predictive models for purchasing in the fashion field by using kernel machine regression equipped with ordinal logistic regression. *Journal of*

Retailing and Consumer Services, *32*, 131–138. https://doi.org/10.1016/j.jretconser.2016.05.008

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

Thomassey, Sébastien, & Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. *Decision Support Systems*, *42*(1), 408–421. https://doi.org/10.1016/j.dss.2005.01.008

Thomassey, Sébastien, & Happiette, M. (2007). A neural clustering and classification system for sales forecasting of new apparel items. *Applied Soft Computing*, *7*(4), 1177–1187. https://doi.org/10.1016/j.asoc.2006.01.005

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324–342.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd edition). Morgan Kaufmann Series.

Wong, W. K., & Guo, Z. X. (2010). A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. *International Journal of Production Economics*, *128*(2), 614–624.

Xia, M., & Wong, W. K. (2014). A seasonal discrete grey forecasting model for fashion retailing. *Knowledge-Based Systems*, *57*, 119–126. https://doi.org/10.1016/j.knosys.2013.12.014

Xu, Y., Dai, Z., Chen, F., Gao, S., Pei, J., & Lai, L. (2015). Deep learning for drug-induced liver injury. *Journal of Chemical Information and Modeling*, *55*(10), 2085–2093.

Yu, Y., Choi, T.-M., & Hui, C.-L. (2011). An intelligent fast sales forecasting model for fashion products. *Expert Systems with Applications*, *38*(6), 7373–7379.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *ArXiv Preprint ArXiv:1212.5701*.

Appendices

Appendix A – Dataset attributes

Table A.1 – Dataset attributes

Attribute name	Category	SS15			SS16		
		Nº prod	Total sales	Mean sales	Nº prod	Total sales	Mean sales
FAMILY	Camurca Verdadeiro	0	-	-	9	19180	2131.1
	Entrelacado	0	-	-	4	13317	3329.3
	Estampado	7	25425	3632.1	8	22924	2865.5
	India	2	5122	2561.0	0	-	-
	Palha	16	63684	3980.3	11	43181	3925.5
	Pastas	17	54824	3224.9	14	69175	4947.1
	Patchwork	14	42895	3063.9	24	10515	4383.1
	Pele Verdadeira	3	6087	2029.0	0	-	-
	Plastico	10	25673	2567.3	1	4226	4226.0
	Praia	10	27107	2710.7	0	-	-
	PVC Basico	54	375992	6962.8	41	418734	10213.0
	PVC Bicho	30	87470	2915.7	18	75728	4207.1
	PVC Estampado	34	156507	4603.1	20	99778	4988.9
	PVC Fantasia	50	151504	3030.1	18	61005	3389.2
	PVC Liso	123	454358	3694.0	125	537976	4303.8
	Tecido Liso	9	22233	2470.3	4	15139	3784.8
	Verniz	0	-	-	3	11886	3962.0
	Vintage	5	18853	3770.6	0	-	-
SUBFAMILY	A4	15	35216	2347.7	4	14384	3596.0
	Bolinha	7	23228	3318.3	2	7600	3800.0
	Falsa	13	42270	3251.5	9	41168	4574.2
	Lancheira	0	-	-	1	2317	2317.0
	Malotes	37	122763	3318.0	44	192810	4382.0
	Mao	2	2743	1371.5	0	-	-
	Mochila	6	18070	3011.7	21	72731	3463.4
	Sacos	61	271285	4447.3	42	249077	5930.4
	Shopper	128	561643	4387.8	75	476795	6357.3
	Tracar	111	427962	3855.5	97	412555	4253.1
	Verdadeira	4	12554	3138.5	5	28007	5601.4
COLOR_TYPE	Cor Unica	277	1128382	4073.6	190	1018805	5362.1
	Multicolor	107	389352	3638.8	110	478639	4351.3
COLOR	Acquamarine	7	35194	5027.7	0	-	-
	Beige	53	206673	3899.5	21	105076	5003.6
	Black	74	313996	4243.2	59	342260	5801.0
	Blue	19	51946	2734.0	6	28252	4708.7
	Blue Jeans	4	8226	2056.5	1	5895	5895.0
	Bright Blue	1	1089	1089.0	0	-	-
	Brown	8	27560	3445.0	5	9815	1963.0
	Burgundy	0	-	-	1	5709	5709.0
	Camel	64	289659	4525.9	51	271044	5314.6
	Coral	14	48729	3480.6	5	24323	4864.6
	Ecru	14	57489	4106.4	13	67266	5174.3
	Fuchsia	2	4616	2308.0	0	-	-

	Golden	3	10157	3385.7	4	7577	1894.3
	Green	11	43217	3928.8	8	27785	3473.1
	Grey	4	15863	3965.8	1	5003	5003.0
	Khaki	0	-	-	1	3828	3828.0
	Light Blue	0	-	-	3	13138	4379.3
	Lilac	1	7107	7107.0	0	-	-
	Lime	1	3436	3436.0	0	-	-
	Mustard	6	20157	3359.5	5	22279	4455.8
	Navy	20	109083	5454.2	26	169800	6530.8
	Orange	9	29660	3295.6	3	11581	3860.3
	Peach	5	10379	2075.8	2	2085	1042.5
	Pink	10	42063	4206.3	14	73897	5278.4
	Red	8	23431	2928.9	21	97822	4658.2
	Skin	5	23410	4682.0	3	9334	3111.3
	Taupe	10	37546	3754.6	14	67881	4848.6
	Turquoise	2	7870	3935.0	0	-	-
	White	23	74989	3260.4	27	105710	3915.2
	Yellow	6	14189	2364.8	6	20084	3347.3
FASHION	Basic	25	111157	4446.3	0	-	-
	Basic Fashion	183	609661	3331.5	124	534110	4307.3
	Distribution Centralized	44	328277	7460.8	41	418734	10213.0
	Trendy	132	468639	35550.3	135	544600	4034.1
SEGMENT	Teenager	189	855661	4527.3	141	836609	5933.4
	Woman	195	662073	3395.2	159	660835	4156.2
STORE TYPE	A	35	47063	1344.7	0	-	-
	BA	28	72178	2577.8	30	66203	2206.7
	CBA	136	510812	3756.0	81	332383	4103.5
	DCBA	185	887681	4798.3	189	1098858	5814.1
PRICE		Min = 14.99 Max = 44.99 Mean = 24.75 St. Dev = 4.0			Min = 19.99 Max = 59.99 Mean = 25.92 St. Dev = 5.2		
SIZE	S	14	46754	3339.6	19	83833	441203
	M	256	917857	3585.4	208	853100	4101.4
	L	109	534986	4908.1	72	555141	7710.3
	LX	5	18137	3627.4	1	5370	5370.0
EXPECTATION LEVEL	M1	45	267598	5946.6	45	252051	5601.1
	M2	246	873477	3550.7	158	708986	4487.3
	M3	76	199929	2630.6	73	200771	2750.3
	SB	17	176730	10395.9	24	335636	13984.8

Appendix B – Evaluation metrics on training dataset

Table A.2 – Evaluation metrics on training dataset

		Regression technique					
		DNN	**DT**	**RF**	**SVR**	**ANN**	**LR**
Evaluation metric	**R^2**	0.875	0.812	0.796	**0.923**	0.807	0.612
	RMSE	820.1	937.4	983.9	**618.5**	1045.9	1696.1
	MAPE	0.147	0.117	0.141	**0.069**	0.208	0.387
	MAE	494.3	434.8	470.7	**253.9**	663.1	1029.2
	MSE	689494	918190	1010160	**393851**	1125462	2911026

Appendix C – Performance of models excluding 'Expectation_level' variable

Table A.3 – Models' performance excluding 'Expectation_level' variable

		Regression technique					
		DNN	**DT**	**RF**	**SVR**	**ANN**	**LR**
Evaluation metric	**R^2**	0.633	0.412	0.423	0.560	0.461	0.405
	RMSE	2265.1	2684.3	2685.2	2453.2	2477.9	3115.5
	MAPE	0.419	0.454	0.444	0.435	0.449	0.472
	MAE	1397.6	1618.1	1611.2	1566.0	1537.8	1756.8
	MSE	5130620	7205361	7210423	6018366	6140134	9706410

BIOGRAPHICAL NOTES

A.L.D. Loureiro is a researcher and PhD candidate in Industrial Engineering and Management from the Faculty of Engineering of the University of Porto. She's been working, for 2 years, in business analytics field and on the application of data mining techniques to develop decision support tools to assist management agents on their decisions and on the enhancement of the business.

V.L. Miguéis is an Assistant Professor in the department of Industrial Engineering and Management at the Faculty of Engineering of the University of Porto, Portugal. She received her PhD in Industrial Engineering and Management from the Faculty of Engineering of the University of Porto. Her research interests include educational mining, customer relationship management, data mining, customer intelligence and forecasting. Her research specifically focuses on the use of data mining techniques to support the decision process. She has published papers in several international journals indexed in the Web of Knowledge. She has taught courses in operations research, data mining, statistics and operations management. She is the external relations manager of the Industrial Engineering and Management Master of the Faculty of Engineering of the University of Porto.

Lucas F.M. da Silva is an Associate Professor with Aggregation at the Department of Mechanical Engineering of the Faculty of Engineering of the University of Porto (FEUP) and Director of the Integrated Master in Mechanical Engineering. He leads the Adhesives Group of FEUP. He is editor in chief of *The Journal of Adhesion* and of *Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications*. He is the chairman of several conferences related to adhesion and materials. He is the president of the Portuguese Adhesion Society. He has over 200 ISI papers and a h-index of 37.

Highlights:

- New fashion product's sales are predicted using data mining regression techniques
- The performance of both deep neural networks and shallow methods is explored
- Expert knowledge is part of the predictive variables of the developed models
- Variables describing physical and distribution characteristics are considered