FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Predicting product sales in fashion retailing: a data analytics approach

Nelson da Silva Alves

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Vera Lucia Miguéis Oliveira e Silva

July 23, 2017

Predicting product sales in fashion retailing: a data analytics approach

Nelson da Silva Alves

Mestrado Integrado em Engenharia Informática e Computação

July 23, 2017

Abstract

In the retail context, an erroneous determination of the amounts to buy of each article from the suppliers, either by excess or defect, can result in unnecessary costs of storage or lost sales, respectively. Both situations should be avoided by companies, which promotes the need to determine purchase quantities efficiently. Currently companies collect huge amounts of data referring to their sales and products' features. In the past, that information was seldom analyzed and integrated in the decision making process. However, the increase of the information processing capacity has promoted the use of data analytics as a means to obtain knowledge and support decision makers in achieving better business outcomes. Therefore, the development of models which use the different factors which influences sales and produces precise predictions of future sales represents a very promising strategy. The results obtained could be very valuable to the companies, as they enable companies to align the amount to buy from the suppliers with the potential sales.

This project aims at exploring the use of data mining techniques to optimize the amounts to buy of each product sold by a fashion retail company. The project results in the development of a model that uses past sales data of the products with similar characteristics to predict the quantity the company will potentially sell from the new products. The project will use as a case study a portuguese fashion retail company which sells women bags. It will also use some text mining techniques to extract data from fashion trends web pages of the next season.

Coefficient of determination (R2) will be used to assess the quality of the model proposed.

Resumo

No mercado de retalho de moda, uma determinação errônea dos montantes a comprar de cada artigo pelos fornecedores, seja por excesso ou defeito, pode resultar em custos desnecessários de armazenamento ou vendas perdidas, respectivamente. Ambas as situações devem ser evitadas pelas empresas, como tal surge a necessidade de determinar as quantidades de compras de uma forma precisa. Atualmente, as empresas recolhem grandes quantidades de dados referentes às suas vendas e características dos seus produtos. No passado, essa informação raramente era analisada e integrada no processo de tomada de decisão. No entanto, o aumento da capacidade de processamento de informações promoveu o uso da análise de dados como meio para obter conhecimento e apoiar os responsáveis pela tomada de decisão com o objetivo de alcançar melhores resultados comerciais. Portanto, o desenvolvimento de modelos que utilizem os diferentes fatores que influenciam as vendas e produzem previsões precisas de vendas futuras representam uma estratégia muito promissora. Os resultados obtidos podem ser muito valiosos para as empresas, pois permitem que as empresas alinhem o valor a comprar aos fornecedores com as vendas potenciais.

Este projeto visa explorar o uso de técnicas de extração de dados para otimizar as quantidades de compra de cada produto vendido por uma empresa de retalho de moda. O projeto resulta no desenvolvimento de um modelo que usa dados de vendas anteriores dos produtos com características semelhantes para prever a quantidade que a empresa venderá potencialmente dos novos produtos. O projeto usará como um caso de estudo uma empresa de retalho de moda portuguesa de carteiras de mulher. Também serão desenvolvidades técnicas de text mining para extrair dados sobre as tendências da moda da próxima estação, a partir de páginas web.

Para validar a qualidade do modelo proposto, serão utilizados o coeficiente de determinação (R2).

Acknowledgements

I would like to thank my supervisor Vera Lucia Miguéis Oliveira e Silva, PhD., for the support, availability and guidance provided throughout this master's degree dissertation.

I would like to thank the Faculdade de Engenharia da Universidade do Porto, which led to the formation of my knowledge in Software Engineering.

I would like to thank all my family, grandparents, godparents, uncles and cousins, for the support and affection that show daily, always available to help in various aspects of my life. To my grandfather Franklim.

I would like to thank my parents and my sister for the support, understanding, patience, dedication, affection and unconditional love in the most difficult hours and at many smiles, joys and fantastic moments.

I would like to thank my friends who helped me, supported me and motivated me to the realization of this project.

To all those who, directly or indirectly, contributed to the this dissertation, and, above all, contributed to my professional growth, thank you!

Nelson da Silva Alves

"There are three methods to gaining wisdom.
The first is reflection, which is the highest.
The second is limitation, which is the easiest.
The third is experience, which is the bitterest. "

Confucius

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
ENN	Evolutionary Neural Networks
ELM	Extreme Learning Machine
EELM	Extended Extreme Learning Machine
GM	Grey Method
RF	Random Forest
SKU	Stock Keeping Unit
SVM	Support Vector Machine

Chapter 1

Introduction

The present dissertation was realized in the scope of Mestrado Integrado em Engenharia Informática e Computação, Faculdade de Engenharia da Universidade do Porto (FEUP). In this chapter, the objectives of the project, the methodology used and the structure of the dissertation will be presented.

1.1 Framework

This project was framed in the area of purchasing management to assist a company in the process of purchasing products for retail. The company's products are part of the women's fashion accessory market. The products to be exhibited to the final customer, are governed by two main collections: Spring-Summer (Spring-Summer) and Autumn-Winter (Fall-Winter). In order to respond to customer needs on time, collection planning begins with the analysis of fashion trends for the next season, by both designers and buyers, in order to define which ones and how many integrate the collection plan. After the arrival of the articles from the collection to the warehouses, these are sent to the stores, according to their needs, defined by their exhibition capacity and their sales flow. Depending on the performance of the products or needs demonstrated by the sales volume in the stores, the design and purchasing teams can develop new products or repeat productions that will enter the market in the next season.

1.2 Problem

Currently, the fashion retail consists of selling new products every season. In general, these products have different characteristics from those introduced in past seasons.

Before reaching the stores the fashion products are passed through a supply chain that begins with the production of products several months before the entrance in the market. This fact imposes that companies forecast the quantity they will sell from each product in order to define the quantity to acquire from the supplier.

An erroneous forecast translates into loss of profit, due to sales losses or due to excess of inventory. Companies are constantly collecting data regarding their sales. Consequently this data may constitute an important source of knowledge to improve the quality of sales forecasts. Indeed, the volume of sales of product from previous seasons, combined with the information of those products characteristics, may give insights into the preferences of the buyer, leading to most accurate sales predictions.

1.3 Objectives

Given the introduced context the objective of this thesis is to develop a prediction model able to estimate the sales of new fashion products according to their characteristics, through the analysis of previous collections sales. It is considered that the sales volume of products from previous collections tend to be similar to the sales volume of products with similar characteristics that belong to the new collection. Each product will be analyzed individually and forecasts will be made for sales of these products. Obtaining correct forecasts will support the members of the business management department in the decision-making process. It will also use some text mining techniques to extract data from fashion trends web pages of the next season.

1.4 Innovation

This topic of modeling from data mining is not new and nowadays is very popular. As such, it is important to highlight what distinguishes this study from previous studies: Innovation goes through the extraction of database knowledge to create a forecast model taking into account: previous sales data and user-generated content. Moreover, regarding content generated by the users, it is understood that this content is potentially useful to forecast sales made available online. There are many social networking users who talk about fashion and tendencies. Therefore, realizing what users are talking about may reflect the attractiveness of the products included in the new collection. Finally, the model will be applied to a real context, through the use of real information from a fashion company.

1.5 Methodology

The execution of the project was divided in several steps, in order to define the problem in a structured way. Thus, the main steps were (seen in figure 1.1):

Preprocess At this stage of the work the dataset that supports the project is analysed. The database includes both training data and test data. The database is studied in order to understand what data it has. Data quality will be analyzed, for example if there is missing data, redundant values, inconsistent information, noisy data, outliers or data with impossible values that need treatment. The data will pass through a statistical analysis, so that it can be better interpreted.

Analysis Model In this phase several models are created , using different approaches to the problem and different resolution strategies, namely, different algorithms. At this stage a method capable of collecting online user-generated data is also developed.

Validate Model At this stage, the models are analyzed and compared to each other. The validation is based on the use of regression measures that will allow us to evaluate the prediction quality against what was expected to happen (eg coefficient of determination (R2)). The test data will be used to evaluate the degree of similarity with the results obtained by the constructed models.

Image

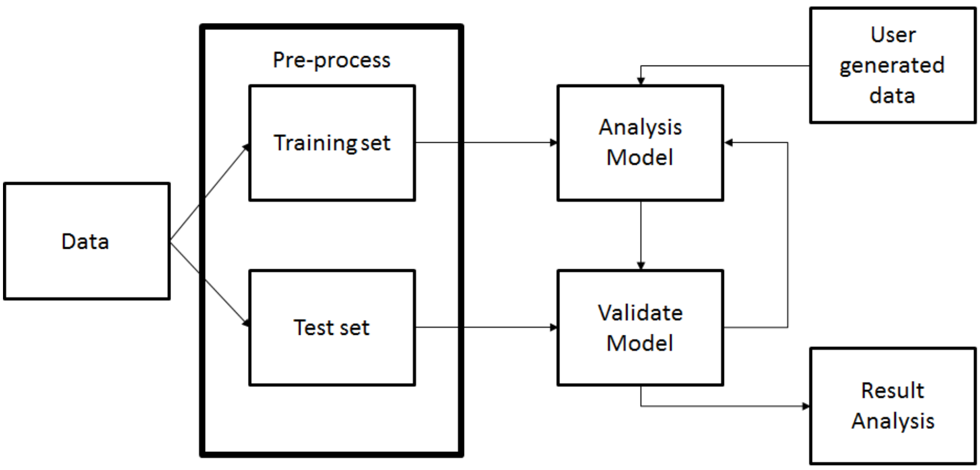

Figure 1.1: Methodology scheme.

1.6 Structure

This report is divided into four chapters.

In the first one, a brief introduction of the problem that motivated this work is presented and the objectives are presented.

In the second chapter, a review of the literature is carried out, explaining the characteristics of this project and the different approaches that other authors have used to solve problems in some way similar to this one.

In the third chapter, the data set is described.

In chapter four, the implementation of the developed program is explained as well as what techniques were used.

Finally, in chapter 5, an evaluation of the performance of the created model is made and the conclusions drawn from the data obtained are presented, as well as the main obstacles encountered.

Chapter 2

Literature Review

2.1 Introduction

Once the problem is related to two different areas, the literature review will be divided into two parts as well: fashion retail and data mining. The first part will describe the main characteristics of the market for which the project will be developed and its influence in forecast models. The second part will describe several methods proposed by different authors to solve sales forecasting problems.

2.2 Fashion Retail

According to a study carried out by Thomassey [1], the fashion forecasting models are influenced by several characteristics. These characteristics will be described throughout the next sections.

In fashion retail, the sale of a product to the customer corresponds to the last step of a complex process by which the product passed. This process corresponds to a chain of steps involving the intervention of several companies. The retail company is the impeller of the chain and is responsible for the sale of the products to the final costumer. As such it is likely that the company foresees the sales that it will make. For a good sales forecast, it is necessary to know first the characteristics of the fashion retail industry:

- Clothing is very much related to the weather making the sales seasonal. Although it is possible to predict the general trends, the different variations of the climate can lead to peaks or hollows.

- Fashion trends provide very volatile consumer demands. The style of the articles should always be updated and the articles are often not repeated for the next collection. Due to the constant novelty and short duration of the articles in the stores, the historical sales are practically non-existent. This leads to a low predictability of future demands.

- Sales are conditioned by many other variables such as end-of-season sale, sales promotion, purchasing power of consumers, etc.
- There are a great variability of products. They can have various colors, shapes and sizes. All of them must satisfy the final consumer.
- In the fashion market, purchases are mostly made by impulse when confronted with the product in a store, not by necessity, so the availability and visibility of the product in the stores is of great relevance, so it is important to have the right product for sale;
- Fashion products have also great instability in demand, as they are usually affected by external conditions such as weather conditions or the use of such articles by celebrities;

Taking all these factors into account, creating a forecast model requires huge knowledge of the subject making it very specific and complex.

2.2.1 Time Horizon

A precision model is primarily based on a sales history associated with a time span of the past. Choosing the right time slot is crucial. It is important to estimate sales using a suitable horizon that is not too large. Accuracy with high anticipation can lead to very high errors. It is necessary to consider the processes associated with the distribution of the products: purchases, orders, replenishments, inventory allocations, etc and taking into account the time associated with: production, shipment, transportation and quality control. Based on this, a horizon of, for example, 1 year becomes adequate. If it is possible to replenish during a sales season, then a horizon of a few weeks may be useful. In the latter case you can also adapt a forecast to analyze the sales of local stores and replenish them if needed. Different horizons involve different methods to compute the forecast model.

2.2.2 Fast Fashion and Product Life Cycle

Since the 1990s, the business model in the clothing industry has changed:

- Styles are now defined based on the interests of the customers rather than of the designers.
- Although collections are still divided basically into two stations, Spring-Summer and Fall-Winter, articles of each collection are and can undergo changes as the season goes by.
- Mass production has reduced due to the change of focus of the companies in the different interests of the costumers.

Based on those characteristics, the standard styles have turned into a huge variety of styles. There are more production of different products in lower quantities, increasing the turnover in stores.

Comparing the article supply network with the product life cycle, it should be noted that in the fashion industry the life cycle is quite short. However, there are still some items that can be sold during all year or during a specific part of the year, as such denims and basic white t-shirt, respectively. Fashion items are the ones that are sold punctually in a short period. Also, there are best selling items which can be sold every year with slightly modifications, based on the fashion trends.

Forecasting the sales of each product becomes a very important task and should take into account the different characteristics of each one. Depending if it is a fashion item or not different approaches of forecasting models should be used.

2.2.3 Seasonality

Another feature that characterizes the fashion industry is seasonality. Every time-series analysis must use the seasonality factor to adjust prediction results. However, in the fashion industry, some items are logically very sensitive to seasonal variation, such as wear swim or pull overs, others are not affected, such as panties. Thus, according to the sensitivity of the item considered, seasonality should be more or less integrated into the clothing sales forecasting system.

2.2.4 Exogenous Variables

The clothing market is heavily impacted by numerous factors that make sales very fluctuating. These factors, also called explanatory variables, are sometimes uncontrolled and even unknown. Some of them involve an increased purchasing decision, others modify store traffic [43]. The impact of these factors can be very different in sales. In fact, some factors generate point fluctuations without significantly affecting total sales volume, for example, the time price discount produces sales peaks. Others impact sales more globally as macroeconomic environment or retail strategy. Therefore, practitioners should keep in mind the following aspects when constructing the forecasting system [59]:

- Explanatory variables are essential to model clothing sales and, if possible, the most relevant should be integrated into the forecasting calculation. The variables are many and varied and it is not possible to establish an exhaustive list

- The impact of each of these variables is particularly difficult to estimate and is not constant over time

Undercontrol	Uncontrolled
Item features and fashion trends	Macro-economic data
Retailing strategy	Calendar data
Marketing Strategy	Competition
	Weather data

Table 2.1: Exogenous Variables.

- These variables can be correlated in them
- Some variables are not available (e.g. competing data) or predictable (e.g. meteorological data) and therefore can not be integrated into the forecasting system.

In table 2.1 it is presented the exogenous variables.

2.2.5 Forecast Errors

The direct effects of forecasting on efficiency, costs, inventory levels or levels of customer service are difficult to understand [4,58]. In the literature, much research has shown that a reduction in prediction errors leads to better supply chain performances [10,29,54,75]. In [34], the authors investigate seven supply chains in different industrial sectors and conclude that a suitable forecasting model allows stabilizing the supply chain, especially for price sensitive products. In [9], an empirical analysis of sales of more than 300 SKUs from a supermarket, clearly shows the relationship between forecasting errors, inventory stocks and inventory costs. In [29], the authors simulate a method to understand and quantify the effect of forecasting on different indicators such as cost, stock level, service level, etc. They find that reducing forecasting errors offers better benefits than choosing inventory decision rules. They also show that an erroneous specification of the forecasting method definitely increases costs. Similarly, [2] investigates the relationship between forecasting and operational performance in the supply chain in the chemical industry. They have shown that choosing the forecasting method strongly impacts customer service and costs. Information sharing, and more especially the sharing of forecasting data, also has a strong impact on supply chain management [3, 15, 42, 73]. According to these studies, it seems obvious that fashion companies have to implement a proper forecasting system and share their forecasts and then try to restructure and / or rethink their supply chain to reduce deadlines and minimum order quantities.

2.3 Sales Forecasting Methods

2.3.1 Pre-processing and Feature Selection

Being the pre-processing an essential step for the data mining was made a small revision on this concept for the project:

- According to Crone et al [65], pre-processing techniques have a major impact on forecasting models. This impact can be both positive and negative.
- Many authors [40,41] compared different models and it was generally agreed that the performance of each model varies significantly according to the level of attributes.
- Given the context of fashion retail, according to King [64] the most relevant predictor factor is color.

2.3.2 Forecasting Methods

As far as forecasting models are concerned, there are a large number of models throughout the literature. In the next sections, the most relevant ones will be analyzed. The models will be categorized in: traditional methods, advanced methods and hybrid methods.

2.3.2.1 Traditional Methods

The use of time series forecasting methods is one of the most commonly used techniques for predicting sales data. These statistical techniques include several models, namely: exponential smoothing [4], Holt Winters [5], Box & Jenkins [6], regression [7]. These methods were implemented in different areas and showed satisfactory results [8]. However, the efficiency of these methods depends on the area to which they are applied, the horizon and even the user experience [9]. Other articles refer other statistical methods such as the extension of standard methods and variants of the Poisson model [10], a model based on the binomial distribution [11], as well as the Croston model and its variants [13] and bootstrap methods [12].

There are also statistical models of analysis of time series such as ARIMA and SARIMA. Since these methods have a closed-form expression for forecasting, it is simple and easy to implement and the results can be calculated very quickly. Another model applied by Green and Harrison [2] uses a Bayesian approach to explore the prediction of a company selling ladies' dresses to order. Another recent work [3] examines the applicability of a Bayesian prediction model to predict fashion demand. It is found that the proposed hierarchical Bayesian approximation produces superior quantitative results compared to many other methods.

Another method is based on a truncated Taylor series [14]. The sales forecast made through a Taylor Series, where the first derivatives are the most important component. The final forecast is calculated from a weighted sum of historical data with more weight for more recent data. In [15], a diffusion model is proposed to predict new product sales. Considering some assumptions, sales are extrapolated from a non-linear symmetric logistic curve considering the saturation level, inflection point and delay factor of the product life cycle.

Although these methods are widely used, especially due to the simplicity and ease of computation they have some disadvantages. It is sometimes difficult to choose the most appropriate statistical method for the forecast in question. These disadvantages go through the difficulty of working on intermittent, erratic or irregular demand data. Traditional prediction methods such as exponential smoothing [34] should be used for smooth, high volume demand and do not work well with intermittent, erratic or irregular demand. These methods are also limited to their linear structure. This type of methods also requires large historical data sets and it is difficult to incorporate other variants such as the exogenous features of the fashion retail market. Thus, pure statistical methods may not achieve a desirable prediction result. Compared to more sophisticated methods, purely statistical methods do not show very promising results. The adoption of other techniques in conjunction with these statistical methods may be one way of overcoming some of these obstacles. Table 2.2 presents some traditional methods.

Forecast Method	Characteristics
Exponential Smoothing Holt Winters Box & Jenkins ARIMA and SARIMA Regression models Poisson model Binomial distribution Croston's model Bootstrap methods	Advantages: • Simple • Compute Results Quickly Disadvantages: • Linearity • Requires a lot of Historical Data

Table 2.2: Some traditional methods.

2.3.2.2 Advanced Methods

According to [16] it is important to use article classification systems to examine the accuracy of predicted sales of new items. They consider that a larger number of item families and relevant classification criteria are required for the respective forecasting procedure in order to obtain better prediction accuracy. They conclude that the product family and aggregate forecast are more accurate than predictions for individual items. Many more advanced and more modern methods use classification techniques for the production of their models.

Artificial neural networks (ANN) are probably the most used techniques for sales forecasting, especially for short-term forecasts, where the main issue is to give more importance to the latest known sales [17]. AI models can handle data with non-linear approximations. The ANN produce good results when the forecast is not seasonal and not fluctuating [18]. For ANN to produce good results it is necessary that they be adapted to the sales forecast, otherwise these techniques become unsuitable for the use in question.

Many authors have obtained quite good results through ANN [19, 20]. Recent studies about artificial neural networks (ANNs) for sales forecasting report their improved performance against more conventional approaches [21].

In the fashion sales forecast literature, Frank et al. [22] explores the use of the ANN model to drive retail fashion forecasting. Comparing it with two other statistical methods in terms of prediction results, it turns out that the ANN model achieves the best performance. Subsequently, the evolutionary neural network model (ENN), which is a promising global approach to selection of features and models, has been used in the fashion sales forecast. To be specific, [23] employ ENN to look for the ideal network structure for a forecasting system and then an ideal neural network structure for the fashion sales forecast is developed.

Despite the fact that the ANN and ENN models present good results in terms of obtaining high prediction accuracy, these techniques also present a disadvantage that may impede their actual application. The disadvantage in neural networks corresponds to the time required to produce the model. Neural networks use gradient-based learning algorithms such as the backpropagation neural network (BPNN). These algorithms are time-consuming and necessary to train the neural network. The model creation time also depends on the complexity and variety of the data used. Being the main function of these models the use for short forecasts and their time-consuming creation of a model becomes a great obstacle to their use.

Recently, extreme learning machine (ELM) algorithms have been extensively described and implemented in the literature for sales forecasting issues, and more especially for the learning process of ANN [24, 25, 26, 27, 28]. Comparing with ANN models with gradient learning algorithms, ELM should be better at generalization and faster at learning [27]. ELM is known to be a fast method and can successfully avoid the problems associated with stopping criteria, learning rate, learning times, local minimums and over-adjustment.

Sun et al. [26] investigate the relationship between the quantity of sales and the significant factors that affect the demand (for example, design factors). Other studies apply the ENN to predict the sales. Performing real data analysis, they show promising results, especially in the case of noisy data [29].

However, ELM has its most critical drawback of being "unstable" because it can generate a the different result in each different run. To overcome this problem, an extended ELM method (EELM) is proposed in [30] which calculates the result of the forecast by repeatedly executing the ELM several times. It is clear that the number of repetitions is an important parameter in the EELM and can be estimated.

Even though ELM and EELM are faster than the classic ANN and ENN based prediction models, they are far from perfect. In particular, EELM still needs a substantial amount of time to perform the prediction. In other words, there are cases where they may not be appropriate [31].

If ELM has demonstrated its effectiveness in the problem of sales forecasting, even in the fashion industry, they may still suffer, as gradient or back propagation methods, from over-fitting or sub-fitting especially to fashion sales data.

The theory of fuzzy sets is proposed by Zadeh [32] and has been applied in many areas. These methods are based on the fuzzy set theory and fuzzy logic. It is commonly used for resolving vague and uncertain information, that are unavoidable in many real-world decision-making processes. In general, uncertain and vague information means that decision making has to be done with relatively unverifiable and inconsistent information without any formal approach. Fuzzy logic and Fuzzy Inference Systems (FIS) are often used to model non-linear, floating, disturbed, and incomplete knowledge and data [33]. These characteristics lead to the implementation of fuzzy inference systems to model complex relationships between data, as an influence of exogenous factors on sales [34]. Comparing with the actual sales models of 322 item families, this system based on fuzzy significantly improves the accuracy of medium term forecast. This result demonstrates that an estimation of some influences of exogenous factors is an important factor to be considered for a sales forecast of fashion items. Sztandera et al. [37] has constructed a new multivariate fuzzy model that is based on many important product characteristics such as color, time and size. On the proposed model, grouped data and sales figures are calculated for each size class. Hui et al. [38]

Forecast Model	Focus
Model using Artificial Neural Networks (C. Frank et al, 2003)	Sufficient data
Model using Fuzzy (L. M. Sztandera, 2014)	Short term, Sufficient data
Model using evolutionary neural network (ENN) (Au, Choi, & Yu, 2008)	Low demand uncertainty and weak seasonal trends, short term
Model using Extreme Learning Machine (Sun, Choi, Au, & Yu, 2008)	Color, size, and price as significant factors
Model using extended EML (Yu, Choi, & Hui, 2012)	Fast forecasting

Table 2.3: Some advanced methods and their focus.

explores the problem of forecasting demand in terms of fashion forecasting. This study uses the fuzzy logic system that integrates the preliminary knowledge of pre-color editing with the fuzzy core prediction system based learning to conduct prediction. They report several promising results of the proposed method.

2.3.2.3 Hybrid Methods

These models are designed to take advantages of different methods at the same time, creating a new model. Due to the use of several techniques in a single model, the statistical models or even as pure ANN end up becoming less efficient. This is well seen in the most recent literature review, where the application of this type of forecasting methods for sales is much studied [39, 40]. Methods used in the fashion forecast literature often combine different models ANN, and ELM with other techniques.

Vroman et al. [41] derived a fuzzy-adaptive model that controls the weighting factors of an exponential-smoothing Holt-Winter statistical prediction method. They prove that the proposed fuzzy hybrid model outperforms the conventional Holt-Winter method. They even argue that the hybrid method can be used for forecasting new fashion item sales. In another model, created by Thomassey and Happiette [35] two automatic systems were combined. In order to deal with the lack of historical data, they propose methods of soft computing: inference systems and neural networks. This approach addresses challenges effectively and has good results [36]. However, they report that such approach can be difficult to be adopted by clothing companies. Another author [42], have applied a hybrid fuzzy model to the fast fashion forecast. They combine the fuzzy logic model with the statistical model to make the forecast. In their approach, they forecast for weekly demand using a weighted average of the predictions generated by many methods. They say that their method is applied very accurately.

In hybrid artificia neural network (ANN) models, ANN can be combined with other techniques such as Grey Method (GM) and autoregressive technique. For instance, Ni and Fan [44] apply a dynamic two-stage prediction model, which contains neural network and auto regression technique, to fashion retail forecasting. In this approach, they use neural networks to establish a multivariable error prediction model. The model develops the concept of "influence factors" and divides "impact factors" into two distinct stages (long and short term). The method results shows that the multivariate error prediction model can produce good forecasting results for fashion retail sales forecasting problems. Aksoy et al. [45] combine the neural networks and the fuzzy method to create a new model called fuzzy inference system based on adaptive network. The proposed model combines the advantages of both techniques, namely the generalization ability of the fuzzy logic technique and the learning ability of neural networks, generating a powerful hybrid model. More recently, Choi et al. [46] applied a GM and ANN based hybrid model to forecast fashion sales with regard to color. They analyze the changing regime of ANN, GM, Markov, and GM + ANN hybrid models. They conclude that the GM and ANN hybrid model is the best to predict color fashion sales when the historical data is small.

The Extreme Learning Machine (ELM) is fast in forecasting [47]. Although not perfect due to its unstable nature, its "fast speed" makes it a very good candidate to be a component model for the most advanced hybrid model for fashion forecasting. For example, Wong and Guo [48] propose a new ANN based on learning algorithms to initially generate the initial sales forecast and then use a fine-tuning heuristic technique to get a more accurate final sales prediction. Its learning algorithm integrates an improved harmony search algorithm and an extreme learning machine to improve network generalization performance. They argue that the performance of the proposed model is superior to the traditional ARIMA models and two models of neural networks recently developed to predict fashion sales. Xia et al. [49] examined a predictive model based on an extreme learning machine model with adaptive metrics. In their model, the inputs can solve the problems of amplitude change and trend determination, which in turn helps to reduce the effect of over-assembly of networks. Yu et al. [50] use Gray relational analysis (GRA) and ELM to create a method of predicting color for the hybrid fashion method. Their model result used real empirical data and has proved that it outperforms several other competing models in predicting fashion color.

In addition to the types of hybrid methods discussed above, there are a few other prediction methods that are also reported in the fashion forecast literature. For example, a hybrid SARIMA wavelet transform (SW) method was employed for predicting fashion sales by Choi et al. [51]. Using real and artificial data, they proved that with a relatively weak seasonality and a great variation of the seaonality factor, the SW method performed better than the classic statistical methods. They said that the proposed method is adequate for the volatile forecast in fashion. Thomassey and Fiordaliso [52] have developed a hybrid method that is based on a decision tree classifier and on an existing clustering technique. The proposed method proved to be good in estimating the sales profiles of new items in fashion retail, when no historical sales data is available. There is another hybrid method proposed by Ni and Fan [53] that establish a combined method that includes the self-regression method and decision tree (called the ART method). They say that the developed hybrid method has a very good performance for predicting fashion sales. Table 2.4 presents some

Forecast Model	Focus
Fuzzy inference systems and neural networks (Thomassey et al, 2002)	Lack of historical data, mean- and short-term forecasting
Model with Fuzzy adaptation of the Holt-Winter (P. Vroman,M.Happiette, and B. Rabenasolo, 1998)	New items
Hybrid intelligent model based on ELM and harmony search algorithm (Wong WK, Guo ZX, 2010)	Mean term
Hybrid SARIMA wavelet transform method (Choi, Yu, & Au, 2011)	Highly volatile sales
Model based on clustering and decision trees (S. Thomassey and A. Fiordaliso, 2006)	Mean term

Table 2.4: Some hybrid methods and its focus.

hybrid methods and their focus.

2.3.3 User generated data

Today, with the popular use of the Internet as a means of communication and information gathering, customers have also started to inform and educate their tastes more deeply. The common user has become an active and productive entity and no longer passive and purely consumer. Customers search for fashion opinions to understand the tendency that they wish to follow in the future. This information is often present on other people's blogs or social networks. Monitoring such information can become an asset to fashion retailers.

According to Kaplan and Haenlein [54] social media is a set of Internet-based applications that allow the creation and exchange of User Generated Content. The Twitter microblogging service has served as the data source for most of the works. For example, Asur and Huberman [55] focus on box receipts and movie data from Twitter and demonstrate high correlations between online data and the actual ranking of the movie. Dhar and Chang [56] suggest that user-generated content is a good predictor of future online music sales. Likewise, Twitter posts were used to examine the role of the platform in predicting the outcome of future elections [57]. Another search flow is the use of search keywords for prediction. Google's influenza trends, for example, estimate flu-based influenza distributions based on influenza-related keywords two weeks faster than another system [58]. They assume a relationship between these keywords and people actually showing flu symptoms [59]. Goel, Hofman, Lahaie, Pennock and Watts [60] have a similar approach: they focus on entertainment products and assume that consumers who are interested in a particular movie or game can look for it online. They report a greater correlation between movie revenue and online activity, as opposed to music-related search queries. Likewise, Kulkarni, Kannan and Moe [61] consider the volume of research as a product interest and a significant indicator for future box office receipts. Beheshti-Kashi and Thoben [63] understand search queries also as a type of user-generated content and thus suggest the combination of both search flows within the user-generated content integration exploration within the fashion forecasting process. However, they propose this approach as a judicious adjustment of baseline forecasts.

Chapter 3

Methodology

The objective of this project is to develop a model for forecasting quantities of fashion products to be sold in the next homologous season. The forecast model will use information about the future trends present in web pages and will also use the data present in the previous year's sales history. For the creation of this model we use techniques of text mining and data mining. Data mining techniques are inserted in linear regression category.

3.1 Text mining

The text mining techniques used will serve to extract useful information from web pages. In this project the technique used initially passes through the selection of words that are related to the training data. The words are grouped and organized according to the demands of the problem in which they are inserted. Once chosen, the words are then used to find the same words on the web page. Different words can sometimes be used to extract data about the same characteristic (e.g. "gray" and "grey"). In this process the frequency of the words is extracted and the data are analyzed and treated. At the end of the process, useful information is obtained that will be used to assist in the forecasting process.

3.2 Data mining: Regression

The linear regression corresponds to the creation of a mathematical model that explains a quantitative output data of an input data set. The obtained mathematical model can then be used for forecasting using different input values resulting in new output values. The regression models are then used to predict actual sales figures such as retail.

3.2.1 Artificial Neural Networks

Artificial Neural Networks (ANN) are distributed systems based on the nervous system and are composed of a set of artificial neurons, constituting processing units. Each artificial neuron has a set of input connections, to receive input values either from an input attribute vector or from other neurons. Each input connection has a weight value associated, simulating the synapses in the nervous system. The network weight values are defined by a learning algorithm. A neuron defines its output value by using an activation function to the weighted sum of its inputs. This output value is sent to the ANN output or to other artificial neurons. Figure 3.1 represents a neural network structure.

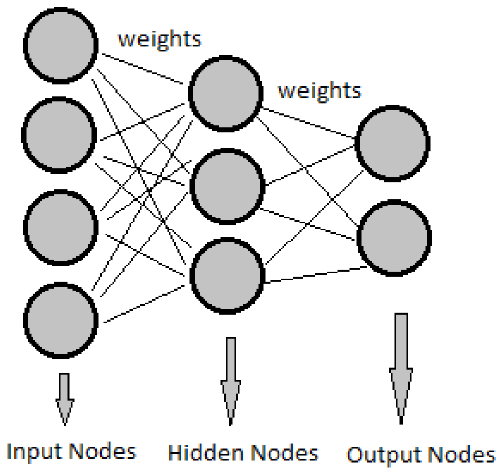

Figure 3.1: Neural Network representation.

3.2.2 Random Forest

Random forests correspond to a combination of several decision trees. Random Forest grow each decision tree using a different bootstrap sample. At each node of the tree, the algorithm only use a pre-defined number of attributes randomly selected.

3.2.3 Support Vector Machine

Support vector machines, SVMs, is a ML technique that reduces the occurrence of overfitting by looking for a model that present high predictive performance and has low complexity. It has a strong mathematical foundation. SVM maximizes the separation margin between the two classes by selecting support vectors among the training objects from the two classes. The position of these support vectors in the input space define the separation margin. The margin of tolerance is called epsilon. The algorithm taken in consideration is based on minimize the error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated. SVM algorithm and formulas are presented in figure 3.2.

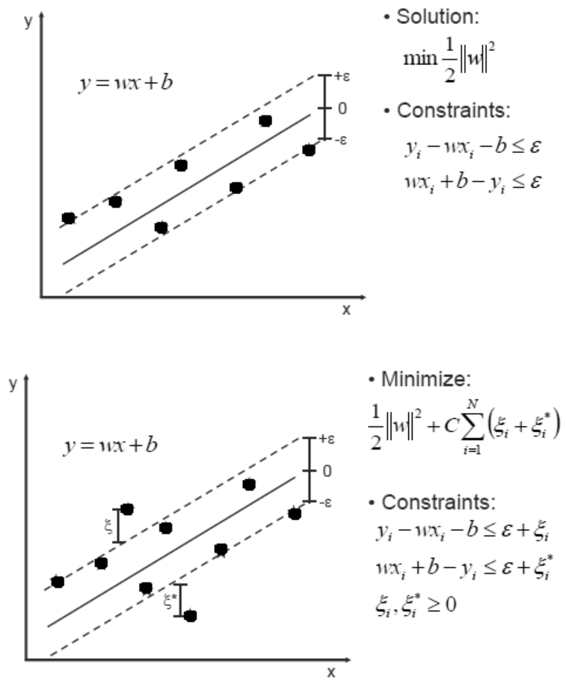

Figure 3.2: Support vector machine representation and its formulas.

The probabilistic regression model used assumes (zero-mean) laplace-distributed errors for the predictions, and estimates the scale parameter using maximum likelihood.

3.3 Model Validation

Finally, for the purpose of measuring the quality of the model, it will be used the values of the forecast quantities and the real values of the quantities sold. As a measure of validation of the model will be used the coefficient of determination (R2). R2 formula is expressed in image 3.3, where SS(regression) corresponds to the sum of squares of the predicted values of an analysis model and SS(total) corresponds to the sum of squares of the real values presented in test data.

$$r^2 = \frac{SS_{(regression)}}{SS_{(total)}}$$

Figure 3.3: Coefficient of determination formula.

Chapter 4

Implementation

The implementation of the project went through several phases namely: data set analysis, preprocessing data set, user-generated data extraction and modeling.

The data set in which this work is based consists of the sales history of handbags sold during the spring/summer season of 2015 and the sales of the homologous period in 2016.

4.1 Data set analysis

Each product entry in the data set has various characteristics which will be described next:

PROD_COD It is the code identification of a product. All entries have different values for this attribute.

SEASON It specifies the season of the product. In the data set used there are only "SS15" or "SS16" products (Spring/Summer 2015 or 2016).

GAMA It is the category of the product. In the data set used there are only "carteiras" (wallets).

FAMILIA It is the family of the product. Values in data set: Beach, True Suede, Varnish, Basic PVC, True Leather, Vintage, Printed PVC, Interlaced, Printed, India, Plain PVC, Briefcase, Animal PVC, Plain, Plastic, PatchworkStraw and Fantasy PVC. Figure 4.1 presents the total sales grouped by family.

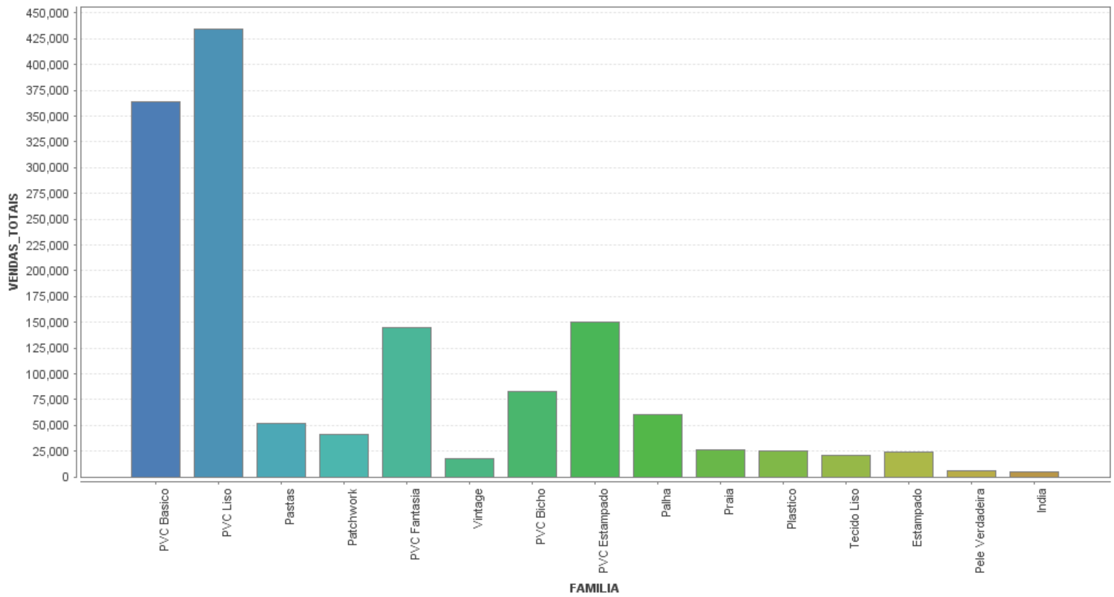

Figure 4.1: Total sales of season Spring/Summer 2015 grouped by family.

SUBFAMILIA It is the subfamily of the product. Values in data set: (Bags), (True), (A4), (Ball), (Shopper), (Interlace), (Lunch Bag), (False), (Backpack), (Hand), (Pouch). Figure 4.2 presents the total sales grouped by subfamily.

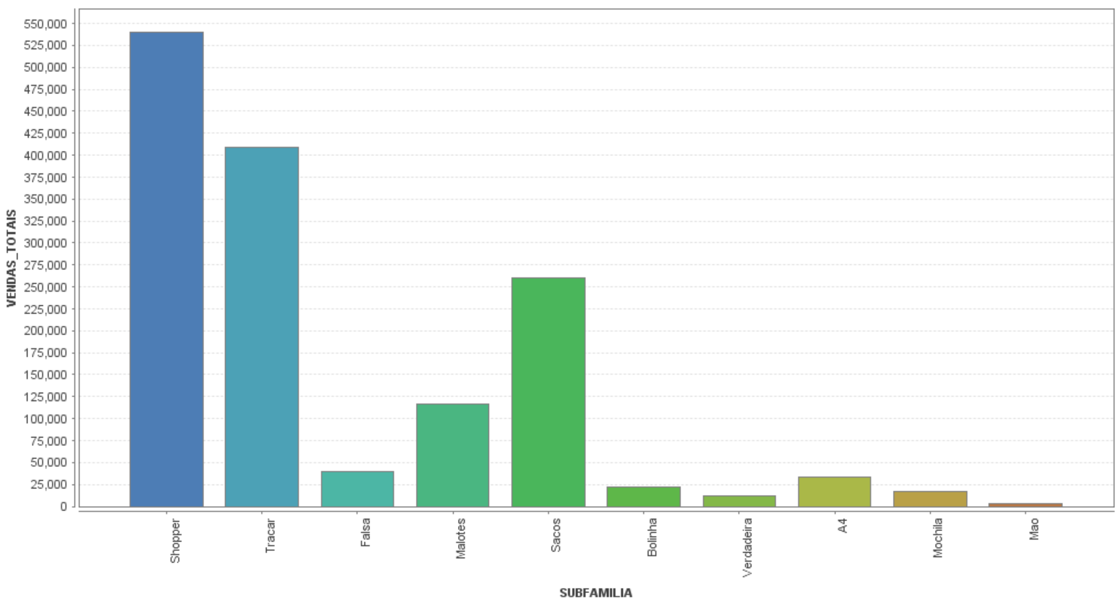

Figure 4.2: Total sales of season Spring/Summer 2015 grouped by subfamily.

TIPO_COR It is the type of color of the product. If a product have clearly a predominant color it has the value "Cor Unica" (unique color), and if a product has various colors its value is

"Multicolor".

COLOR It is the predominant color of the product. It can be "Golden", "Mustard", "White", "Pink", "Fuschia", "Coral", "Navy", "Ecru", "Skin", "Yellow", "Green", "Grey", "Brown", "Blue", "Turquoise", "Beige", "Camel", "Black", "Orange", "Lilac", "Burgundy", "Lime", "Blue Jeans", "Khaki", "Peach", "Acquamarine", "Red", "Bright Blue", "Taupe" and "Light Blue". Figure 4.1 presents the total sales grouped by color.

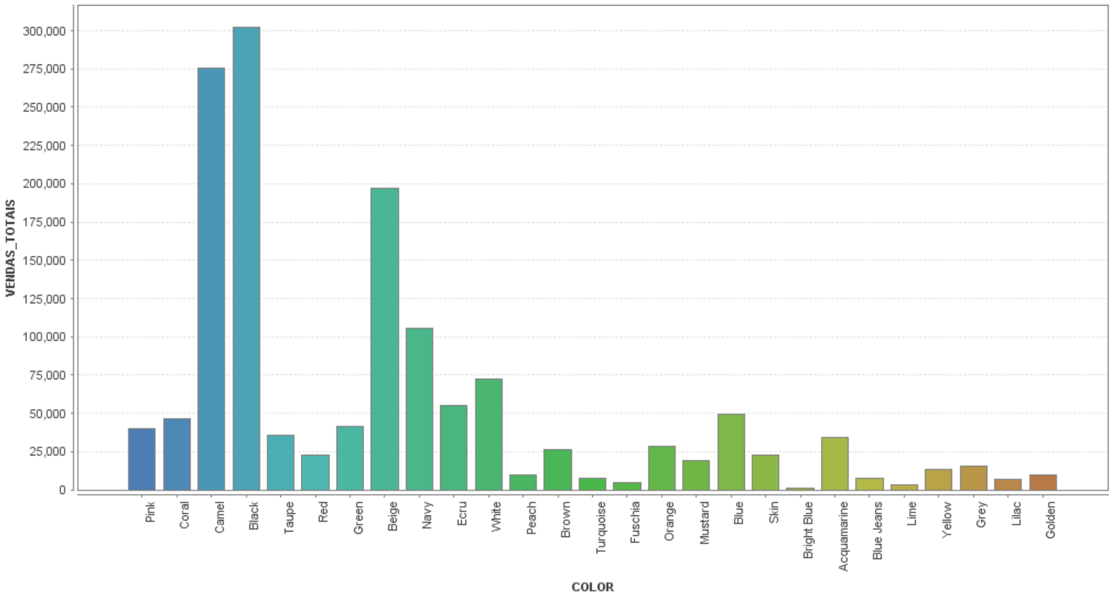

Figure 4.3: Total sales of season Spring/Summer 2015 grouped by color.

FASHION It is associated with the tendencies of the season, being an article classified as "Moda" if this is considered to be part of the tendencies of the season, followed by "Básico Moda" and, finally, "Básico". Cases classified as "Centralized Distribution" are articles similar to others existing in past seasons and whose behavior is predicted to be similar because they are alike. Depending on the fashion type it is assumed that the article is displayed in store a different numbers of weeks. It can be "Básico" (basic - eight weeks), "Básico Moda" (Basic Fashion - six weeks), "Distribuição Centralizada" (centralized distribution - eight weeks), "Moda" (fashion - four weeks).

INTERNACIONAL Feature that defines whether the item may or may not go to all markets. Articles classified as "N" are articles that can go to all markets, the remaining cases are specific.

RESPONSÁVEL It is the creator of the product. There are "Marta Fragateiro", "Nídia", "Teresa" and "MJM/Elena".

MATCHING They are articles that do "Matching" with other gammas. It can be "S" or "N".

SEGMENTO It is the segment of the product. It can be "Teen" or "Woman".

TIPOLOGIA It corresponds to the classification given to the store where the item is sold. It can be "A", "BA", "CBA", and "DCBA". It represents the designated ratio of typology (A, B, C or D) depending on the relation between the store size (Large, Normal, Average, Small) and its sales potential (1, 2, 3, 4 or 5), seen in table 4.1.

Size/Sales Potential	1	2	3	4	5
Big	A	A	A	B	C
Normal	A	A	B	B	C
Medium	A	A	B	C	C
Small	B	B	C	C	D

Table 4.1: Relation between store size and sales potential.

A large and potential selling shop 1 will be considered a typology A store, which will receive all of the items in a collection, while a small and potential selling store 5 will receive a smaller variety of the items in the collection because it is a typology store D. This classification by typologies is used as an aid in determining the minimum quantities of each article to be purchased, and this value is obtained by aggregating the quantities to be sent to the stores associated with a certain typology. As the purchase decision of the articles is based on the type of stores in which it will be exposed, agglomerations of typologies were defined to help calculate the purchase quantities. For example, stores defined as type D in a given range are stores whose sales expectations of this range are low and therefore there are no items that are purchased exclusively for this typology. This means that stores with type D will receive only products that have been purchased to supply the entire store universe, making the "Typology" of these items defined as "DCBA". On the other hand, stores classified as A can, without problems, receive specific articles because they have the capacity to sell them, that is, the characteristic "Typology" will be defined as A.

PREÇO_BASE_IVA It is the price of the product in euros. It is a numeric value.

TAMANHO It is the size of the product. In the data set there are "S", "M", "L" and "XL".

APOSTA It is associated with the sales prediction of the company department for the product and its average sales per store. Values in dataset: "M1", "M2", "M3" and "SB".

VENDAS_TOTAIS It is the total sales of the product at the end of the season.

	PROD_COD	SEASON	PROD_COD_1	GAMA	FAMILIA	SUBFAMILIA	FAMILIA_EQ	SUBFAM_EQ	TIPO_COR	COLOR
1	130622RO_L	SS15	130622RO_L	Cart...	PVC B...	Shopper	Shopper	PVC Basico	Cor Unica	Pink
2	130792CL_M	SS15	130792CL_M	Cart...	PVC Liso	Tracar	Tracar	PVC Liso	Cor Unica	Coral
3	130831CA_XL	SS15	130831CA_XL	Cart...	Pastas	Falsa	Falsa	Pastas	Cor Unica	Camel
4	130835CA_M	SS15	130835CA_M	Cart...	Patch...	Tracar	Tracar	Patchwork	Cor Unica	Camel

Figure 4.4: Data set example part 1

FASHION	INTERNACIONAL	RESPONSÁVEL	MATCHING	SEGMENTO	TIPOLOGIA	PREÇO_BASE_IVA	TAMANHO	APOSTA	VENDAS_TOTAIS
Distrib...	ND	Teresa	N	Teen	DCBA	22,99	L	SB	10152
Moda	ND	Teresa	N	Teen	DCBA	22,99	M	M2	3706
Moda	ND	Teresa	N	Teen	CBA	29,99	XL	M2	4063
Moda	ND	Teresa	N	Teen	CBA	24,99	M	M2	3516

Figure 4.5: Data set example part 2

Comparing both periods it should be noted that the items sold in the 2015 and 2016 seasons are different, however some have similar characteristics. Figure 4.4 and 4.5 presents a sample of the data set used.

4.2 Preprocessing data set

The database includes both training and testing data. Initially, the data were separated according to the two stations in which they were inserted. Divided by attribute SEASONNAME. The Spring Summer 2015 data contains the training data and the data for Spring Summer 2016 constitute the test data. Analyzing the database, it was noticed that some tuples did not have values related to total sales. These data were filtered and removed from the data set as they did not have a relevant value for predicting the data. Once the price attribute is a numerical value it was decided to normalize it using the Z-transformation method. No redundant values, inconsistent information, noisy data, no outliers or data with impossible values were found. No other preprocessing methods were made. VENDAS_TOTAIS is the attribute models will predict in Spring-Summer season 2016.

4.3 User-generated data extraction

A user-generated data extraction program was developed by the author in order to extract potentially useful information from web pages. The web pages to choose from should contain useful information relative to the characteristics of what will be fashionable for the intended season. As such, web pages were chosen that predicted the characteristics for the seasons under study. For Spring/Summer 2015 season the web pages that were chosen preceded the year of 2015 and for the Spring/Summer 2016 season were chosen web pages that preceded 2016. Five web pages were used to extract information related to the SS 2016 season. As regards the season of 2015, it was more difficult to find web pages with potentially useful content and therefore 10 web pages were used. The text that is present on the web page is collected and filtered. If the words found in the web page matches the words presented in the regular expression then that information is collected. The filtered words are previously selected and are related to the products' characteristics in the database. For this project were chosen words referring to the color, family and subfamily of the products. For the characteristics present in the database words were chosen to be counted from the web pages. They are listed in tables 4.2, 4.3, 4.4, 4.5, 4.6:

Color related words															
Dataset	Golden	Mustard	White	Pink	Fuschia	Coral	Navy	Ecru	Skin	Yellow	Green	Grey	Brown	Blue	Turquoise
Searched	golden	mustard	white	pink	fushia	coral	navy	ecru	skin	yellow	green	grey, gray	brown	blue	turquoise

Table 4.2: Color related words used to search on web pages.

Color related words															
Dataset	Beige	Camel	Black	Orange	Lilac	Burgundy	Lime	Blue Jeans	Khaki	Peach	Acquamarine	Red	Bright Blue	Taupe	Light Blue
Searched	beige	camel	black	orange	lilac	burgundy	lime	jeans	khaki	peach	acquamarine	red	bright blue	taupe	light blue

Table 4.3: More color related words used to search on web pages.

Family related words									
Dataset	Praia	Camurca Verdadeiro	Verniz	PVC Basico	Pele Verdadeira	Vintage	PVC Estampado	Entrelacado	Estampado
Searched	beach	suede	varnish	basic	leather	vintage	printed Pvc	interlaced	printed

Table 4.4: Family related words used to search on web pages.

Family related words									
Dataset	India	PVC Liso	Pastas	PVC Bicho	Tecido Liso	Plastico	Patchwork	Palha	PVC Fantasia
Searched	india	plain Pvc	briefcase	animal	plain	plastic	patchwork	straw	fantasy

Table 4.5: More family related words used to search on web pages.

Subfamily related words											
Dataset	Sacos	Verdadeira	A4	Bolinha	Shopper	Trancar	Lancheira	Falsa	Mochila	Mao	Malotes
Searched	bags	true	A4	ball	shopper	interlaced	lunch bag	false	backpack	hand	pouch

Table 4.6: Subfamily related words used to search on web pages.

The frequency of each word present on each page is summed and grouped by word. At the end of this process, you get the sum of the absolute frequencies of each word. At the end, the relative frequency is calculated and this information is conditioned to the database.

4.4 Data Junction

The previously calculated information is then included in the database. Three new attributes are created, namely: "COLOR_FREQ", "FAMILY_FREQ" and "SUBFAMILY_FREQ". The relative frequency of each characteristic is associated with the corresponding products characteristics, can be seen in figure 4.6.

PROD_COD	VENDAS_TO...	PREÇO_BAS...	SUBFAMILY...	FAMILY_FREQ	COLOR_FREQ	APOSTA	COLOR
138018MZ_L	11184	-1.137	88.889	0	1.087	SB	Navy
138643RO_S	7174	-0.562	0	0	4.348	M1	Pink
138660BE_L	4574	0.779	0	6.061	0	M2	Beige
139904CL_M	5550	1.354	0	0	0	M1	Coral
138018BE_L	13233	-1.137	88.889	0	0	SB	Beige
139051CA_M	4838	-0.562	0	0	0	M2	Camel
139634BE_S	4759	-1.137	0	0	0	M2	Beige
139634MZ_S	4747	-1.137	0	0	1.087	M1	Navy
139677BE_M	4226	-0.179	0	0	0	M2	Beige
139875PR1M	4464	0.779	1.111	6.061	14.130	M2	Black
140321CA_M	5259	-0.562	0	0	0	M2	Camel
140322BR_M	4504	-0.179	0	0	25	M2	White
139729TP_M	2436	-0.179	0	0	0	M2	Taupe
140296PR1M	5201	-1.137	0	0	14.130	M2	Black
140297VE_M	3163	-0.179	88.889	0	7.609	M3	Red
139698CR_M	169	-0.562	0	0	0	M2	Ecru

Figure 4.6: Data set after user-generated information has been included.

4.5 Analysis Model

After the data was collected and the test and training data sets were properly treated, it was decided to use three different methods to predict the amount of sales. These methods were chosen because they were presented among the most used methods for prediction problems, as analyzed in the literature review chapter.

The three models that will be used for tests are:

Random Forest - Uses several sets of decision trees for forecasting. The use of a greater number of decision trees in the process is generally associated with more accurate results.

Support Vector Machine - This model allows the regularization of some parameters and the type of kernel to be used, allowing the user to have some flexibility in the way he trains the data.

Neural Network - It is a model inspired by the central nervous system of an animal have the advantage of capturing and dealing well with the existence of errors in the data.

Chapter 5

Results

5.1 Results Obtained

The results obtained for three different models using support vector machine, random forest and neural network are presented in Table 5.1. In figure 5.1 it is possible to compare the difference between real sales values (blue points), random forest forecast values (orange points) and support vector machine forecast values (gray points) of Spring-Summer season 2016 sales.

For the following models, techniques were used to tune and optimize the parameters involved in the corresponding models. In this way the presented values correspond to the best values in each model of several executions.

The measure used to evaluate the performance of the models was the coefficient of determination between the predicted sales figures and the actual values of the test data of the following homologous season. In the measurement used, the performance is evaluated on a scale between 0 and 1, where values close to 0 correspond to forecasts of quantities of sales that are far from the correct values and a value of 1 corresponds to the correct forecast of the quantities of items to be sold.

	Coefficient of Determination (R2)	
	Without user-generated data	With User-generated data
Random Forest	**0.8079**	**0.8127**
Support Vector Machine	**0.6939**	**0.6976**
Neural Networks	**0.7391**	**0.6144**

Table 5.1: Models performance.

Of the three models used, the one that produced closer estimates to the real sales was the model that used Random Forest. This model predicted with a coefficient of determination of 0.8079 for the data set that did not include user-generated data. With the junction of user-generated data the forecast was 0.8127. With the other models the prediction with user-generated data was lower, reaching 0.6976 using a Support Vector Machine approach and the one which used Neural Networks had reached 0.6144 of coefficient of determination.

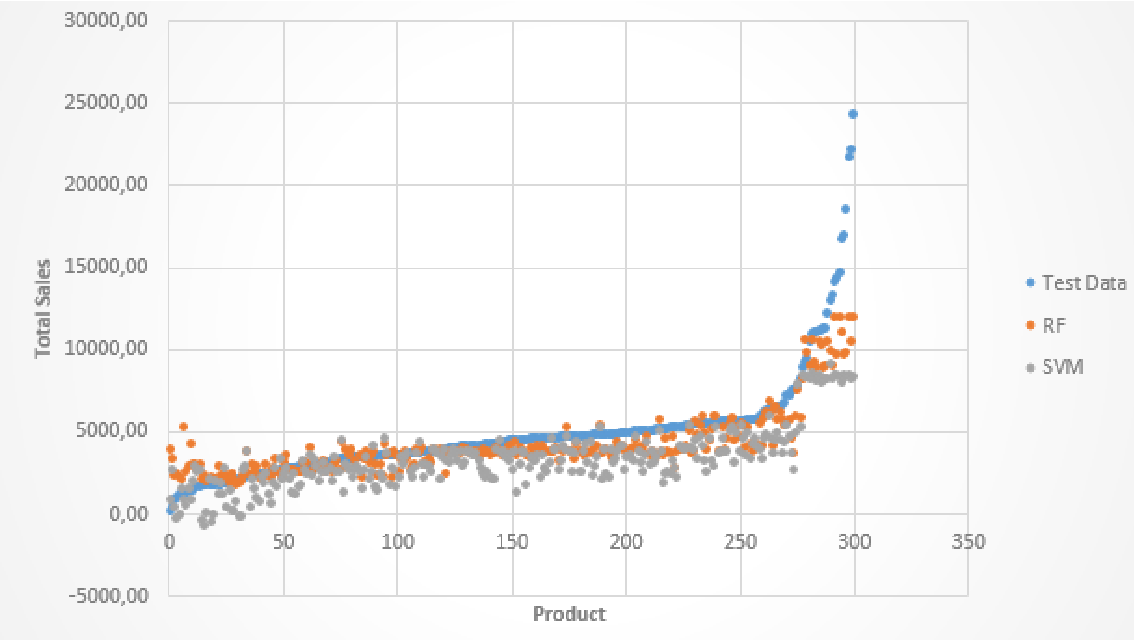

Figure 5.1: Models performance.

Figure 5.1 shows the products that were used for the forecast and their respective sales quantities. Each number on the horizontal axis represents a random product with specific characteristics that are different from the other product. The blue points represent the actual sales of the spring summer 2016 season items and are distributed in the chart in ascending order of sales. The predictions of the Random Forest and Support Vector Machine models are represented by the orange and gray points, respectively. By the analysis of the graph, it should be noted that the forecasts using the Random Forest model are closer to the real sales than the forecasts obtained by the Support Vector Machine model. It should be noted that in the articles where sales are higher, the forecasts of the models present greater discrepancy to the real values.

```
only 20 most important variables shown (out of 81)

                                     Overall
TIPOLOGIACBA                          100.00
TIPOLOGIADCBA                          68.62
APOSTAM3                               65.56
TIPOLOGIABA                            57.38
APOSTAM2                               54.95
APOSTASB                               47.39
FASHIONDistribuição Centralizada       30.22
COLORCamel                             28.21
INTERNACIONALVERMELHOS                 27.88
FAMILIAPlastico                        25.17
FAMILIAPVC Basico                      24.52
FAMILIAPraia                           24.30
PREÇO_BASE_IVA                         24.01
FAMILIAPVC Estampado                   23.48
FAMILIAPVC Bicho                       22.78
FASHIONModa                            21.59
FASHIONBásico Moda                     21.30
COLORRed                               21.12
SUBFAMILY_FREQ                         20.63
FAMILIAPVC Liso                        20.07
```

Figure 5.2: Importance of variable of a Random Forest forecast.

In Figure 5.2 the importance of the variable is shown. The preponderance of the "Tipologia" and "Aposta" attributes were, initially, expected because both allow a more specific categorization of the article regarding sales when compared to the other attributes.

It should also be noted the presence of the variable SUBFAMILY_FREQ that arises with an importance of 20.63 and whose value is very similar to several other variables, not highlighting as being one of the variables with the most impact on the forecast. The variable FAMILY_FREQ and COLOR_FREQ do not appear in the list of the 20 most important variables.

Chapter 6

Conclusions and Further Work

6.1 Conclusions

As mentioned throughout this dissertation, the objectives were to create a model that uses information available on the Internet in conjunction with a set of data to predict the quantities of items to be purchased for the next homologous season. The data set used was related to the historical sales of women wallets for the Spring/Summer season of 2015 and 2016, containing more than 1000 different product entries with 17 different attributes related to their characteristics and total sales.

Using the characteristics of the articles and the information extracted from the internet, a prediction model supported by a data mining technique was created in order to predict the purchase quantities of articles for the new station. Three different regression methods were used to create the model, among which Random Forest was the one that produced the best results. The results show that the proposed model presents a predictive capacity whose coefficient of determination is around 80,79%. With the addition of new information present on the internet, the predictive capacity of the model increased to 81,27%. The insertion of the information present in the web pages was favorable for the forecast of future sales in the chosen model, however, the impact was small. The low impact can be related to the small sample of online content.

Despite the result obtained, it should be noted that the information available online reveals great potential for forecasting future trends, as analyzed in the literature review.

Created the model, this can thus be used to estimate the quantities to buy for the homologous season of the following year.

6.2 Future Work

The future work for the developed project relies mainly on the exploitation of the online content and the text mining techniques for extracting online data but the exploration of the models can also reveal some positive results.

An example of exploration that can be performed is the analysis of the terms used to search by categories or characteristics of the products of the database. This textual exploration applied to the contents available in the Internet can result in a positive impact for the forecast. It may also be useful to look for information from other online sources such as social networks.

The exploration of new forecasting models can also help to have a better outcome, but once this issue has been well worked out and analyzed throughout this project, this task reveals to have less potential than the text mining techniques and online content exploitation.

References

1. Thomassey, S. (2014). Intelligent fashion forecasting systems: Models and applications. In T.-M. Choi, C.-L. Hui, Y. Yu(eds.), pp. 9–27.

2. M. Green and P. J. Harrison, "Fashion forecasting for a mail order company using a bayesian approach," Operational Research Quarterly, vol. 24, no. 2, pp. 193–205, 1973.

3. P. M. Yelland and X. J. Dong, "Forecasting demand for fashion goods: a hierarchical Bayesian approach," in Handbook on Intelligent Fashion Forecasting Systems, 2013.

4. Brown RG (1959) Smoothing forecasting and prediction of discrete time series. Prentice Hall, Englewood Cliffs

5. Winters PR (1960) Forecasting sales by exponential weighed moving averages. Manag Sci6:324–342

6. Box GEP, Jenkins GM (1969) Time series analysis forecasting and control. Prentice Hall, Englewood Cliffs

7. Papalexopoulos AD, Hesterberg TC (1990) A regression-based approach to short-term system load forecasting. IEEE Trans Power Syst 5:1535–1547

8. Kuo RJ, Xue KC (1999) Fuzzy neural networks with application to sales forecasting. Fuzzy Sets Syst 108:123–143

9. Armstrong JS (2001) Principles of forecasting – a handbook for researchers and practitioners. Kluwer, Norwell

10. Wang H.J., Chien C., Liu C., (2005). Demand Forecasting Using Bayesian Experiment with Nonhomogenous Poisson Process Model. International Journal of Operations Research, 2 (1), pp. 21-29.

11. Cachon G., Fisher M., (2000). Supply chain inventory management and the value of shared information, Management Science, 46(8), pp. 1032–1048.

12. Varghese V., Rossetti M.D., (2008). A Parametric Bootstrapping Approach to Forecast Intermittent Demand. Industrial Engineering Research Conference Proceedings, May 17-21, 2008, Vancouver, Canada.

13. Croston J.D., (1972). Forecasting and stock control for intermittent demands. Operational Research Quarterly 23(3), pp. 289–303.

14. McLean, BL,Wortham AW(1988) Methodology aids forecasting with limited amounts of data. Ind Eng 20(2):18–22

15. Morrison J (1996) How to use diffusion models in new product forecasting. J Bus Forecast Methods Syst 15(2):6–9

16. S. Thomassey, M. Happiette, and J. M. Castelain, "Mean-term textile sales forecasting using families and items classification," Studies in Informatics and Control, vol. 12, no. 1, pp. 41–52, 2003.

17. Yoo H, Pimmel RL (1999) Short-term load forecasting using a self-supervised adaptive neural network. IEEE Trans Power Syst 14(2):779–784

18. Wong WK, Guo ZX (2010) A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. Int J Prod Econ 128(2):614–624

19. Chang P.C., Wang Y.W., Liu C.H., (2007). The development of a weighted evolving fuzzy neural network for PCB sales forecasting. Expert Systems with Applications, 32(1), pp. 86 96.

20. Ling S.H., (2010). Genetic Algorithm and Variable Neural Networks: Theory and Application, Lambert Academic Publishing, German.

21. 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.

22. C. Frank, A. Garg, A. Raheja, and L. Sztandera, "Forecasting women's apparel sales using mathematical modeling," International Journal of Clothing Science and Technology, vol. 15, no. 2, pp. 107–125, 2003.

23. K.-F. Au, T.-M. Choi, and Y. Yu, "Fashion retail forecasting by evolutionary neural networks," International Journal of Production Economics, vol. 114, no. 2, pp. 615–630, 2008.

24. Chen FL, Ou TY (2011) Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry. Expert Syst Appl 38:1336–1345

25. Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: Proceedings of the international joint conference on neural networks (IJCNN2004), Budapest, pp 25–29

26. Sun ZL, Choi TM, Au KF, Yu Y (2008) Sales forecasting using extreme learning machine with applications in fashion retailing. Decis Support Syst 46:411–419

27. Wong WK, Guo ZX (2010) A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. Int J Prod Econ 128(2):614–624

28. Zhu Q, Qin A, Suganthan P, Huang G (2005) Evolutionary extreme learning machine. Pattern Recognit 38(10):1759–1763

29. 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.

30. Y. Yu, T.-M. Choi, and C.-L. Hui, "An intelligent quick prediction algorithm with applications in industrial control and loading problems," IEEE Transactions on Automation Science and Engineering, vol. 9, no. 2, pp. 276–287, 2012.

31. T.-M. Choi, C.-L. Hui, and Y. Yu, "Intelligent time series fast forecasting for fashion sales: a research agenda," in Proceedings of the International Conference onMachine Learning and Cybernetics (ICMLC '11), pp. 1010–1014, Guilin, China, July 2011.

32. L. A. Zadeh, "Fuzzy sets," Information and Computation, vol. 8, pp. 338–353, 1965.

33. Zadeh LA (1994) Soft computing and fuzzy logic. IEEE Soft 11(6):48–56

34. Kuo RJ (2001) A sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. Eur J Oper Res 129:496–517

35. Thomassey, S., Happiette, M. (2007). A neural clustering and classification system for sales forecasting of new apparel items. Applied Soft Computing, 7(4), 1177–1187.

36. 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.

37. L. M. Sztandera, C. Frank, and B. Vemulapali, "Predicting women's apparel sales by soft computing," in Proceedings of the 7th International Conference on Artificial Intelligence and Soft Computing (ICAISC '04), pp. 1193–1198, Zakopane, Poland, June 2004.

38. C.-L. Hui, T.-W. Lau, S.-F. Ng, and C.-C. Chan, "Learningbased fuzzy colour prediction system for more effective apparel design," International Journal of Clothing Science and Technology, vol. 17, no. 5, pp. 335–348, 2005.

39. Q. Wu, "The hybrid forecasting model based on chaotic mapping, genetic algorithm and support vector machine," Expert Systems with Applications, vol. 37, no. 2, pp. 1776–1783, 2010.

40. W. I. Lee, B. Y. Shih, and C. Y. Chen, "A hybrid artificial intelligence sales? Forecasting system in the convenience store industry," Human Factors and Ergonomics in Manufacturing Service Industries, vol. 22, no. 3, pp. 188–196, 2012.

41. P. Vroman,M.Happiette, and B. Rabenasolo, "Fuzzy adaptation of the Holt-Winter model for textile sales-forecasting," Journal of the Textile Institute, vol. 89, no. 1, pp. 78–89, 1998.

42. E. Yesil, M. Kaya, and S. Siradag, "Fuzzy forecast combiner design for fast fashion demand forecasting," in Proceedings of the IEEE International Symposium in Innovations in Intelligent Systems and Applications (INISTA '12), pp. 1–5, 2012.

43. P. Vroman, M. Happiette, and C. Vasseur, "A hybrid neural model for mean-term sales forecasting of textile items," Studies in Informatics and Control, vol. 10, no. 2, pp. 149–168, 2001.

44. Y. Ni and F. Fan, "A two-stage dynamic sales forecasting model for the fashion retail," Expert Systems with Applications, vol. 38, no. 3, pp. 1529–1536, 2011.

45. A. Aksoy, N. Ozturk, and E. Sucky, "A decision support system for demand forecasting in the clothing industry," International Journal of Clothing Science and Technology, vol. 24, no. 4, pp. 221–236, 2012.

46. T.-M. Choi, C.-L. Hui, S.-F. Ng, and Y. Yu, "Color trend forecasting of fashionable products with very few historical data," IEEE Transactions on Systems, Man and Cybernetics Part C, vol. 42, no. 6, pp. 1003–1010, 2012.

47. Y. Yu, T.-M. Choi, and C.-L. Hui, "An intelligent fast sales forecasting model for fashion products," Expert Systems with Applications, vol. 38, no. 6, pp. 7373–7379, 2011.

48. W. K. Wong and Z. X. Guo, “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, vol. 128, no. 2, pp. 614–624, 2010.

49. M. Xia, Y. C. Zhang, L. G.Weng, and X. L. Ye, “Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs,” Knowledge-Based Systems, vol. 36, pp. 253– 259, 2012.

50. Y. Yu, C.-L. Hui, and T.-M. Choi, “An empirical study of intelligent expert systems on forecasting of fashion color trend,” Expert Systems with Applications, vol. 39, no. 4, pp. 4383–4389, 2012.

51. T.-M. Choi, Y. Yu, and K.-F. Au, “A hybrid SARIMA wavelet transform method for sales forecasting,” Decision Support Systems, vol. 51, no. 1, pp. 130–140, 2011.

52. S. Thomassey and A. Fiordaliso, “A hybrid sales forecasting system based on clustering and decision trees,” Decision Support Systems, vol. 42, no. 1, pp. 408–421, 2006.

53. Y. Ni and F. Fan, “A two-stage dynamic sales forecasting model for the fashion retail,” Expert Systems with Applications, vol. 38, no. 3, pp. 1529–1536, 2011.

54. Kaplan, A. M., Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. Business Horizons, 53(1), 59–68.

55. Asur, S., Huberman, B. A. (2010). Predicting the future with social media. 2010 IEEE/WIC/ACM International

56. Dhar, V., Chang, E. a. (2009). Does chatter matter? The impact of user-generated content on music sales. Journal of Interactive Marketing, 23(4), 300–307.

57. Bermingham, A., Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results. In Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the international joint conference for natural language processing (IJCNLP), 13 November 2011, Chiang Mai, Thailand.

58. Boulos, M. N. K., Sanfilippo, A. P., Corley, C. D., Wheeler, S. (2010). Social web mining and exploitation for serious applications: Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. Computer Methods and Programs in Biomedicine, 100(1), 16–23.

59. Google. (2014). Google Trends.

60. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., Watts, D. J. (2010). Predicting consumer behavior with Web search. Proceedings of the National Academy of Sciences of the United States of America, 107(41), 17486–17490.

61. Kulkarni, G., Kannan, P. K., Moe, W. (2012). Using online search data to forecast new product sales. Decision Support Systems, 52(3), 604–611.

62. Eysenbach, G. (2006). Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. AMIA . . . Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, pp. 244–8.

63. Beheshti-Kashi, S., Thoben, K.-D. (2014). The usage of social media text data for the demand forecasting in the fashion industry fashion sales forecasting and the predictive power of

online. In Dynamics in logistics. Fourth international conference, LDIC 2014 Bremen, Germany, February 2014 Proceedings, pp. 1–4.

64. King, J. A. (2012). Color trends and selection in fashion design. Colour Design. Theories and Applications. A volume in Woodhead Publishing Series in Textiles, pp. 535–550.

65. Sven F. Crone, Stefan Lessmann, Robert Stahlbock. Impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. In European Journal of Operational Research, 173 (2006) 781–800.