Item attribute weight in fashion assortment planning

Ana Rita Amorim de Araújo Novo

Master's Dissertation

FEUP Supervisor: Prof. Paulo Osswald

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Integrated Master's in Industrial Engineering and Management

2019-06-30

Abstract

Fashion industry is a market in constant change, which leads to high unpredictability of next season sales. Future season's collections have to be defined and ordered with at least one season of anticipation, making the planning of future collections a very sensitive issue for a company's success. Said success is very dependent of the forecasting ability of the company to predict, among other things, the assortment of product mix with relevant attributes to answer customer demand. In this context, the assignment of weights to those attributes constitutes a number of criteria that helps guide retailers selection of product variety for the following collection.

This project proposes a methodology to improve one of the steps taken in the assortment planning process, describing how it is performed in one of the biggest retail database and IT solutions applications. The goal is to provide a selection of product attribute weights that defines how products shall be selected for the next collection, while achieving a given managerial goal (in this case, maximization of sales).

A visual analysis was performed with historical data, aiming at finding relationships when crossing different inputs that could be reflected on sales. Afterwards, three different algorithms (Random Forest, Artificial Neural Network and Gradient Boost) were constructed and tested in order to identify the best-fitting methodology. In all cases, the goal was to extract features importance, in order to be able to assign them a priority order.

The methodologies were applied to large data from a multibrand Mexican fashion company and three different scenarios were defined, differing on the output variable to predict. The ordered scenarios predicted sales units, weekly sales units and a defined ratio, respectively.

It was observed that Neural Networks algorithm would not converge, probably due to the fact that all input features were categorical. A better performance was observed for Gradient Boost over Random Forest. Considering the order of magnitude from output variables in the three scenarios, the predictive error was considered to be high. Percentages were estimated for different feature weights when building the models and an example was provided allying their priority order and knowledge extracted from visual analysis. The presented methodology is flexible to the aggregation of product more convenient for the user and so is the choice of scenario to use.

Ponderação de atributos de produto no planeamento de composição da coleção na indústria da moda

Resumo

A indústria da moda é um mercado em constante mudança, o que leva a uma elevada imprevisibilidade das vendas de coleções futuras. A definição da próxima coleção deve ser efetuada e encomendada com pelo menos uma estação de antecedência, o que leva a que o bom desempenho da empresa se torne fortemente suscetível ao planeamento de sucessivas coleções. Sucesso tal mostra-se muito dependente da capacidade de previsão da empresa, relativo, entre outras coisas, à seleção de produto com atributos relevantes para satisfazer a procura do cliente. É neste contexto que surge a ponderação dos pesos aos atributos de produto como forma de criar critérios de escolha que guiam os retalhistas na composição das coleções

Este projeto propõe uma metodologia para melhorar um dos passos que constituem o processo de planeamento de composição de produto, descrevendo como decorre numa das aplicações de grandes bases de dados e soluções de tecnologias de informação. O objetivo é providenciar pesos de atributos das peças de vestuário que definam a forma como deviam ser escolhidos para próxima coleção, otimizando um determinado objetivo de gestão, neste caso, a maximização de vendas.

Para esse fim, foi realizada uma análise visual com dados históricos, procurando encontrar correlações entre atributos que poderiam ser refletidas nas vendas. Foram posteriormente contruídos e testados três algoritmos diferentes (*Random Forest*, *Artifitial Neural Networks* e *Gradient Boost*) para identificar a metodologia mais adequada. Em todos os casos, o objetivo era extrair a importância das variáveis preditivas, de tal forma que permitisse atribuir-lhes um escalonamento de prioridades.

As metodologias foram aplicadas a um grande conjunto de dados de uma empresa de roupa multimarca mexicana, onde foram definidos três cenários diferentes para análise, distinguindo-se pela variável de saída. Os cenários ordenadamente previam vendas unitárias, vendas unitárias semanais e o e um rácio criado.

Observou-se que as redes neuronais não convergiram, provavelmente devido ao caráter categórico de todas as variáveis preditivas. Entre os restantes algoritmos, o *Gradient Boost* apresentou um melhor desempenho. Considerando a ordem de grandeza das variáveis de saída nos três cenários, o erro associado à previsão foi considerado elevado. Foram estimadas percentagens para os pesos das variáveis preditivas aquando da construção dos modelos, sendo posteriormente apresentado um exemplo que aliava a sua ordem de prioridade com as conclusões retiradas da análise gráfica. A metodologia apresentada é flexível ao nível de agregação de produto que for mais conveniente ao utilizador bem como a escolha do cenário a utilizar.

Acknowledgements

I would like to thank firstly to all people from Retail Consult – KSR, S.A. who gave me the opportunity to collaborate with them, sharing their experience and providing a very friendly work environment. A greater thank you goes to the planning team, specially to the those who were always available for clearing any doubts, Carlos Lima, Paulo Marques and Nuno Santos. Within the company and the University as well, I would like to thank Ana Andrade and Maria João Robles who were not only great team mates but also shared their knowledge and opinions when needed, becoming a great company during the past few months.

A big thanks goes also to Professor Vera Miguéis who despite having a busy agenda, was tireless on guiding me and giving me her best advice. I would like to thank Professor Paulo Osswald who always tried to be as fast and efficient as possible in orienting my dissertation.

I want to thank all my friends who supported me when work was harder or seemed to have no solution. People who made me believe I could take the project to the end and never gave up on my capacities to face this challenge. A big thanks to Daniela Pinho who besides her friendship, shared her wise and brilliant ideas to help me reach a more efficient way of solving the proposed problem, from an innovative perspectve. A big thanks to Sofia Casanova and Veniamin Dyachenko whose presence, words, and capacity of listen were of great importance in a phase of stress and hard work.

Last but not least, I would like to thank my brother and sister whose presence always cheered me up and made me feel home no matter the place I was. A big thanks to my mother for being so supportive and always taking the best care of me. And finally to the one person who allowed me to get here, making the biggest of sacrifices to guarantee my success and happiness, my beloved father. Thank you.

Contents

Acronyms and Symbols

AI – Artificial Intelligence

ANN – Artificial neural networks

ARIMA – Autoregressive integrated moving average

CB – Catboost

ELM – Extreme learning machine

IT – Information Technology

MAE – Mean absolute error

MAPE – Mean absolute percentage error

MSE – Mean squared error

PoC – Point of Commerce

RF – Random Forest

RMSE – Root mean squared error

SARIMA – Seasonal autoregressive integrated moving average

SKU – Stock keeping unit

List of Figures

List of Tables

1 Introduction

1.1 Retail Consult – KSR, S.A. - Overview

Retail Consult provides IT consultancy services to companies in the retail industry. The company is a gold partner of Oracle, a company that provides software for several retail business areas in order to support strategic and operational decisions (Consult 2015b).

Retail Consult's specialists take care of the implementation, customization and support of the functionalities of Oracle software. Their main goal is "Retail transformation, covering Merchandising, Planning & Optimization, Supply Chain, Store and Warehouse Operations, Integration and Architecture" (Consult 2015a). The company has several teams allocated to client's projects, covering different areas shown in Figure 1.

The company was founded in 2011 by 6 entrepreneurs. In 8 years, it has grown exponentially, reaching 300 collaborators, and spread to 6 countries – Portugal, Germany, U.S.A., Mexico, Chile and Brazil. Currently, it has 20 customers whose business fits in one of the following industries: grocery, fashion, pharmacy, do-it-yourself, electronics and telecommunications (Consult 2019).

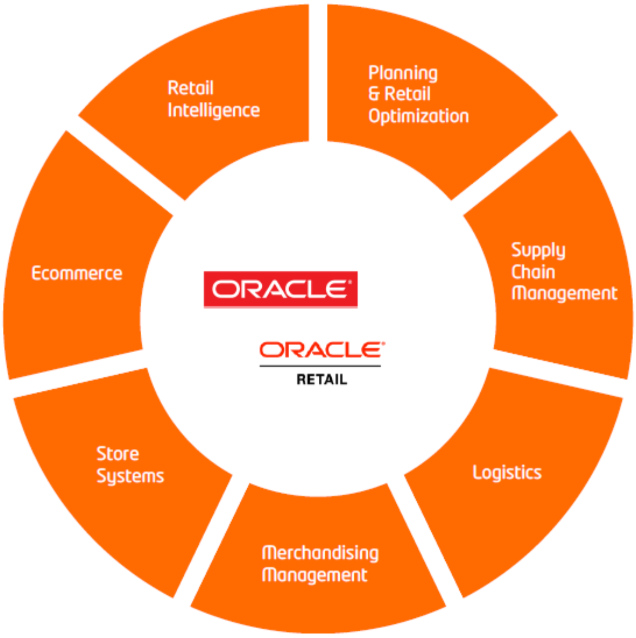


Figure 1 – Retail Consult's services (Consult 2019)

1.2 Motivation

Assortment Planning is a complex, stepwise process that fashion companies need to do as they are preparing the collections to come, in which they assort items to certain Points of Commerce (PoC), in specific periods of time. A Point of Commerce (PoC) is a distribution channel, such as physical stores or e-commerce. Currently, empirical methods are used for the assortment, where decisions are based on users' expertise and choice and on their experience along with analyzing historical sales of different items. This leads to decisions highly biased on user's experience. Thus, a new challenge is unleashed to find other ways to support this kind of decision in a more sustained way.

The motivation of this project is to explore some new opportunities to answer this challenge that comes up in a process that is transversal to all fashion retailers. Thus, this project searched and tested a solution that added data-based logical tools to improve decisions so that they no longer depend on the final user's intuition only. This proposal of study came up in one of KSR's Planning and Optimizing team's internal projects and its findings could be widely applied in several software applications that are currently in the market.

1.3 Objectives

The goal of the project is to overcome the challenge of depending only of experience for the identification of the key attributes that will drive the assortment decision, by defining a model or methodology that determines the best combination of item attributes for fashion retailers to achieve a defined managerial goal, such as sales volume maximization, for instance. The study of item attributes weights is performed in order to improve this empirical decision system, so that it has some data foundation for the decisions taken and no longer depends only on specialist's subjective opinion.

The question being answered by this dissertation is the identification of one of many factors that can impact fashion retail and understand their relative weight, in concrete cases. Thus the purpose of the project rests on learning more about this market's opportunities, helping companies to perform better in a very competitive market.

1.4 Methodology

The methodology approached to answer the mentioned objectives rests on three major steps: data preprocessing, visually analyzing historical data and machine learning models building. After these three phases, the alliance of knowledge extracted from the second and third steps is used for criteria selection.

Data from a fashion company was provided in order to explore the method proposed. Firstly this data required preprocessing which was a step-wise process, consisting on:

- Building a database to structure information;
- Detecting and removing irrelevant information;
- Defining the aggregation level towards time, space, product and consumer;
- Analyzing outliers;
- Transforming features;
- Defining output variables to predict.

Regarding the visual analysis, an overview was done on how each individual variable would impact sales and, afterwards, correlating two and sometimes three variables to look for relationships between features. For this, Tableau software was used to facilitate the graphic representation. This approach was chosen given the fact that all variables except for the target variable were categorical, which made it harder to look for correlations in a non-visual way. All graphs obtained showed sales distribution related to other variables, instead of its sum or average. This decision was taken so that a better comprehension could be made from how data was spread.

For building the models, firstly research had to be done in order to find the best packages from R and Python that could provide not only the predictions but also features importance from training the model. Afterwards, for each model, parameters for tuning were selected and models were built in three phases – parameter tuning, training the model and testing its performance. After finishing this phase, it was possible to extract attributes importance from each model.

Lastly, an example of the application of the alliance between information from the last two steps is exposed, trying to depict in a real case application the solution proposed.

1.5 Dissertation structure

This document is organized in five chapters, where Chapter 2 explores literature review, focusing mainly on forecasting techniques applied to fashion, and providing an overview of the fashion industry characterization, the predictive methods explored in the area and an explanation of the algorithms chosen along with the performance evaluation measures. Chapter 3 defines the problem, by giving an overview of the assortment planning process and providing a concrete example of a widely used application that is inserted into the assortment planning software tools. Chapter 4 explains the methodology approached to reach the proposed solution being divided in preprocessing, visual analysis and data mining models construction. It also presents models results, exploring their application in a concrete example. Chapter 5 draws some conclusions about this work.

2 Review of forecasting techniques applied to fashion industry

Considering the problem proposed, the review was directed at finding a better way to understand the customer's shopping behavior and fashion industry characteristics so that attributes can be weighted according to demand and effective sales. Exploring forecasting techniques was understood as the best way to extract this kind of knowledge, considering the lack of linearity and the variability that make fashion sales behavior so particular comparing to other retail areas.

2.1 Fashion industry

The Fashion industry currently has a very competitive environment, facing demanding challenges such as globalization and uncertain demand (Thomassey and Happiette 2007). Thus, it is required that companies have a well-defined and controlled supply chain, so that lead times of products are taken into consideration in the planning process (Thomassey and Fiordaliso 2006). In order to differentiate their companies from the competition and improve customer's satisfaction, retailers must "provide right products at the right time while maintaining a good stock" (Xia et al. 2012, 253).

Many sales forecasting methods have been studied, in order to help predict the best possible assortment for retailers to answer customer's wants and needs. However, apparel items have many factors that influence their demand and are not easily predictable. First of all, every season, new fashion tendencies are defined, and therefore, new items are produced. Thus, there is no available direct historical data to rely on, as products are always changing (Thomassey and Fiordaliso 2006). Furthermore, customer's taste and choice priorities, such as item's durability, comfort, style, quality and even ethical issues are factors that will impact client's decision too (Loureiro, Miguéis, and da Silva 2018; Jegethesan, K.; Sneddon, J. N.; Soutar 2007).

Having a good sales forecast can help retailers achieve more accurate information regarding products to order. Companies can then take better advantage of their resources achieving more effective ways of optimizing them and reducing costs. A good prediction can also allow to minimize the number of lost sales and provides the possibility of achieving higher profits (Thomassey 2010).

Item and customer's characteristics both influence demand. However, weather conditions, promotions, holidays and economic factors will play an important role as well (Loureiro, Miguéis, and da Silva 2018). Most retailers provide a high variety of Stock Keeping Units (SKU) that will go through a short life cycle on the respective PoC (Loureiro, Miguéis, and da Silva 2018). This particular feature needs to be considered since products are not present during the whole season, but only for a few weeks. As products are arriving to complete the collection, others are leaving. This dynamism is essential to attract customers because it enables them to discover new items every time they visit the PoC, however it increases the complexity of the forecasting process.

For all these motives, demand forecast plays a prominent role on the fashion field. However, few solutions have been found that help retailers understand which of the apparel items' attributes (for instance color or fabric) are going to meet customer's preferences and thus, be translated into sales.

2.2 Current methodologies

This section presents the mathematical and heuristic methods that are currently used for sales forecast, and review examples where these methods and their variations were applied and the respective results obtained.

2.2.1 Statistical and heuristic methods for prediction

Considering all the conditions mentioned about the fashion industry, several sales forecasting models have been studied over the years. These methods can be classified into statistical or mathematical methods and modern heuristic methods (Xia et al. 2012). Regarding the first ones, which mainly deal with time series, the most commonly used are ARIMA, SARIMA, exponential smoothing, regression models, Box and Jenkins, Holt Winters (Loureiro, Miguéis, and da Silva 2018; Thomassey and Happiette 2007), trend extrapolation and seasonal index (Ni and Fan 2011).

Despite the high variety of methods available, and their satisfactory performance in many other fields (Thomassey 2010), when applied in fashion retail there are several drawbacks, due to the assumptions required from the database. To be applied they need a dataset large enough to allow for patterns, trends and cyclical behavior to be found (Choi et al. 2012).

Another relevant aspect relates to the assumption of linearity, which restricts some of these models very much (Xia et al. 2012). Fashion items sales suffer from a strong influence of uncontrollable factors such as volatile fashion tendency, weather events and market price changes (Wong and Guo 2010). Thus, it is very unlikely for linear trends and repetitive cycles to be found (Wong and Guo 2010). Besides that, item are replaced by new ones, every season, so that very little continuous historical data exists on the product (Thomassey and Happiette 2007). Because of this, the analyst has often to aggregate data at higher level of customer segment, product classification or time, since the PoC SKU level does not allow him to have enough information for a good analysis. Along with the transformation of qualitative into quantitative data, these issues turn up as disadvantages for the use of this kind of forecasting techniques (Loureiro, Miguéis, and da Silva 2018). The uncertainty and volatility intrinsic to the market lead to an increase of complexity when optimizing the models parameters (Thomassey 2010).

On the other hand, heuristic methods are being increasingly adopted to perform forecast in fashion. There are several soft computing techniques that have been applied in order to improve decision support systems, namely in the assortment planning process (Ibrahim 2016). The textile industry shows fluctuating, disturbed and incomplete (Thomassey 2010) data, and because of that, the use of data mining techniques tries to tackle the unsolved problems of time series models, through the use of Artificial Intelligence. Fuzzy methods, neural networks, evolutionary algorithms (Thomassey 2010) and expert systems (Wong and Guo 2010) are the main tendencies of AI and they have been showing very positive results when it comes to identifying irregular patterns. These robust and imprecision-tolerant techniques (Ibrahim 2016) are more flexible and thus can adapt better to the peculiarities of the data.

However, according to the chosen model, some handicaps may become evident, regarding problems such as hard interpretability, long running time, and overfitting to the data. Keeping this in mind, studies have focused into exploring hybrid models that combine the advantages of time series and data mining techniques.

2.2.2 Application of prediction methods on real cases

Xia et al. (2012) applied a hybrid method combining extreme learning machine (ELM) and adaptive metrics to forecast sales, obtaining a better accuracy than artificial neural networks (ANN), auto-regression and extreme learning machine alone. The main difference between ELM and ANN relies on the definition of the output weights and biases, which in ELM are randomly assigned, leading to a higher learning speed and a higher generalization performance.

For predicting the future color trend, a comparison of several models was made in a Choi et al. (2012) paper including ANN, a Grey Model, a hybrid between the two and a Markov switching approach. It was concluded that the hybrid model has the best accuracy but that it ran slow, even with a small dataset. The second best model was ANN, even though it was the slowest among all.

Another method was proposed for sales forecasting where a Harmony Search approach and ELM were put together and applied to aggregated data. The aggregation was done by city and category, for different time scales: month, quarter and year. The model was generally good, but, as the aggregation level increased, its accuracy reduces (Wong and Guo 2010).

One more comparison of models was performed, this time between ANN, the fuzzy logic approach and a hybrid method including both of them. ANN showed to be very dependent on the data and the results showed that the hybrid method was no better than the individual methods (Thomassey 2010).

Despite all the existing literature on sales forecasting in fashion industry, little research has been done to try to measure the impact of item attributes on the final sales. However, there were some exceptions, such as Loureiro, Miguéis and da Silva which performed a comparison of shallow data mining techniques and deep learning neural networks. Afterwards, a sensitivity analysis was performed to understand the relative importance of the explanatory variables used in the model (2018). The deep learning method provided results which were hard to interpret and was shown to be a computationally expensive strategy, even though it could also be applied to small datasets. On the other side, methods like Random Forest and ANN had good performances as well (Loureiro, Miguéis, and da Silva 2018). Figure 2 shows the steps taken in model comparison approach.

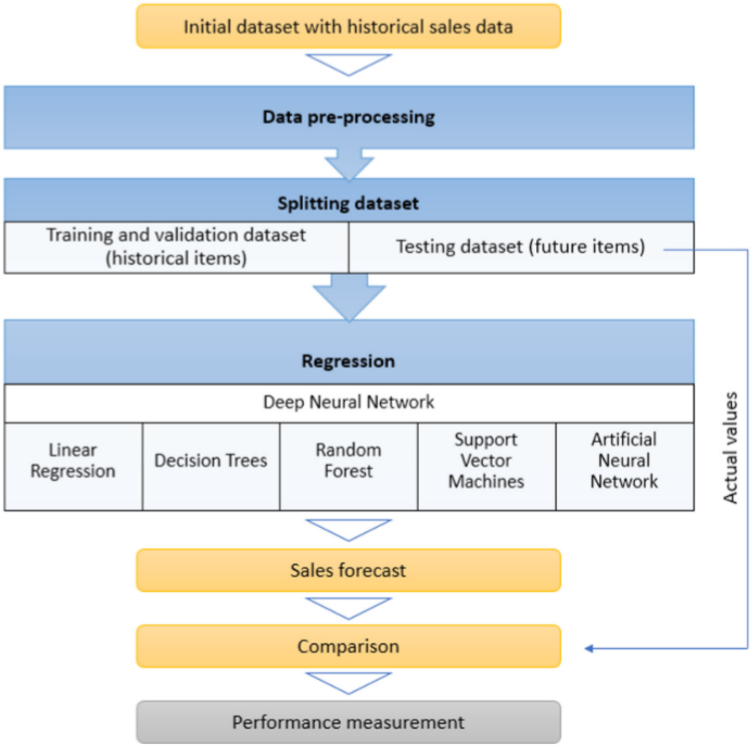


Figure 2 – Methodology taken for model comparison in (Loureiro, Miguéis, and da Silva 2018)

There are two other studies that combine clustering with classification methods. In both situations, sales data was provided, and several product profiles were created through clustering methods, grouping items according to their sales curve. In the first paper, by Thomassey and Fiordaliso (2006) a decision tree was constructed to understand how lifespan, price and starting time of sales influenced each of these product profiles. Afterwards, Thomassey and Happiette (2007)chose ANN as classification method. These classification techniques are usefull because they receive information from new items and can link them to a given sales pattern from one of the profiles. The same explanatory variables were used in both studies. A disadvantage of the decision tree method related to the computational time required for the iterative process (Thomassey and Happiette 2007; Thomassey and Fiordaliso 2006).

Furthermore, some research has been done in order to include human judgement in the forecasting models and understand its impact on prediction. Baecke, De Baets, and Vanderheyden (2017) studied two diferent ways of doing this, either by using a restrictive judgement model or an integrative judgement model. The first one constrains the forecast according to an expert's opinion. The second includes that kind of information in the model as a predictive variable. Human judgement input can add great value to the forecast since experts can identify exceptional events that algorithms could not be trained to identify and predict (Baecke, De Baets, and Vanderheyden 2017). On Loureiro, Miguéis, and da Silva's (2018) paper, domain experts' opinion was included in the predictive models as a variable and showed up to be the most relevant variable to explain the sales, followed by products' physical attributes.

It is important to understand that, according to the kind of study the analyst wishes to perform, data is handled differently. Some studies look only at sales along time, aggregating PoC or product categories in a given time scale (Ni and Fan 2011; Shoham 2008). In other situations, item behavior is evaluated according to its lifecycle, distinguishing basic items, from fashion items and bestselling items, for instance (Thomassey 2010). In this case, basic item are the ones that are sold no matter what time of the year it is, they are permanent in the assortment. Fashion item have short lifecycles and are produced only once. The bestselling items are slightly changed and reappear every year. Wong and Guo (2010) decided to study medium-term sales, aggregating data at category level and city for several time periods. A distinction between high-end, medium and basic fashion item was also performed (Xia et al. 2012). Few studies exist on SKU level in a sense that the products' physical attributes are usually not taken into consideration. So, for the current study, the category aggregation, time scale and PoC group will also be taken into consideration, according to the data and goal to achieve.

2.3 Data mining algorithms

This section is dedicated to the explanation of the theoretical concepts behind the data mining techniques explored in the project.

Data analysis is a science that makes use of distinctive approaches and analyses different types of data, by applying supervised or unsupervised learning methods. Unsupervised learning methods are used when data does not have a target attribute or a specific output, and this is where clustering is included. As for supervised learning, it looks for patterns to reach a given target variable. Supervised learning is usually performed in one of two distinct ways, classification and prediction, also mentioned as regression. Both supervised learning methods share the common goal of obtaining information out of a dataset, through the production of models, differing only on the output they provide. Classification models result in categorical labels. Prediction models, also known as regression models, provide continuous-valued functions (Han and Kamber 2001). This project applies regression models, in order to predict sales, and further analyzes their sensitivity to the item attributes .

2.3.1 Random Forest

For review of the RF method, the book *Data Mining: Concepts and Techniques* from Han and Kamber (2001) was used as reference.

Random Forest cannot be explained without understanding the way decision trees proceed. For that reason, a brief foundation on Decision Trees is provided bellow.

"A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions." (Han and Kamber 2001, 284) An example of the structure described is presented in a schematic way in Figure 3.

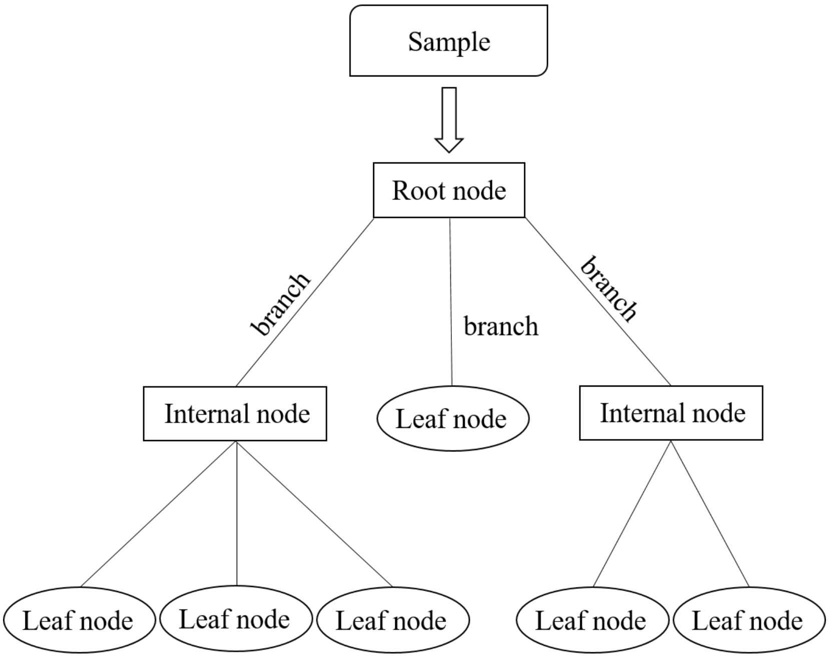


Figure 3 – Schematic representation of decision tree structure based on (Han and Kamber 2001)

The decision tree predicts a classification by making it choose a path in the tree structure, where in every node a test regarding an attribute is answered defining the branch that will lead it to the next node. When it reaches a leaf node, the observation's outcome is predicted.

A training sample is used for creating the decision tree model, where the selected attributes first follow a heuristic that finds the ones which separate the data into distinct subsets more clearly according to an attribute selection measure. When the data in a subset all belong to a common class, the internal node (the attribute associated to it), becomes a leaf node instead, so there is no need to apply the heuristic mentioned any further. For every distinctive value an attribute can take, a branch is formed, and that is where samples will be partitioned. This process stops when it reaches one of three conditions: the whole group of observations from the subset belongs to a common class; or all attribute values have already been used and therefore the highest frequency value from the attribute appearing on the subset becomes a leaf node; or no observations have a given attribute value and a leaf node is created with the majority of attribute values of that subset.

The attribute selection measure is usually information gain or entropy reduction. The information gain measure aims to minimize the amount of information required to classify observations into partitions, making them as little random as possible and it is described through Equation (2.1). It is possible to see how to compute entropy through Equation (2.2).

$$I(s_1,s_2,\ldots,s_m) =- \sum_{i=1}^{m} p_i \log_2 p_i \qquad (2.1)$$

Where:

I, is the information needed to classify a training set S
s_i, is the number of samples that belong to class of object i
p_i, is the probability of a sample belonging to class of object i

$$E(A)=\sum_{j=1}^{v}\frac{s_{1j}+\ldots+s_{mj}}{s}I(s_{1j},\ldots,s_{mj}) \qquad (2.2)$$

Where:

E(A), is the entropy/expected information based on partitioning into subsets by A
A, is the attribute tested for splitting (with j=1 to v attribute values)
s_{1i}, is the number of samples that belong to subset j and have attribute A

The analyst is looking forward to having the minimum entropy possible in all partitions since it is a measure of impurity. Equation (2.3) shows the entropy reduction associated to knowing the attribute A. The goal is to get the highest possible information gain.

$$\text{Gain}(A) = I(s_1,\ldots,s_m)-E(A) \qquad (2.3)$$

Where:

I, is the information needed to classify a training set S
E(A), is the entropy/expected information based on partitioning into subsets by A

A common problem that may come up in several data mining techniques happens when the model obtained includes outlying and irregular patterns from the training observations which are not representative of the population. This problem is defined as overfitting and in order to handle that situation in decision trees, tree pruning is usually performed. Tree pruning aims to remove least reliable branches and it can be done during the creation of the decision tree or after it is formed. The first case is named prepruning and it implies that some branches are transformed into leaf nodes instead of being further partitioned. It usually works with a chosen threshold so there is a required minimum number of partitions. The second case is postpruning and it works by calculating the error of branches after having a fully-grown tree and cutting them off according to the adopted postprunning approach. The decision can be taken by minimizing error or computational cost, according to the analysis conditions and requirements.

The big advantages of decision trees rely on their very user-friendly and intuitive interpretability and their capacity to deal with non-linear patterns. However, some disadvantages inherent to their nature come up, such as the lack of capacity of handling outliers correctly which leads to producing overfitting models. With the purpose of improving these problems and still making use of its advantages, random forests were put into practice.

A random forest is an ensemble method which aggregates the results from several randomly chosen decision trees, each of them grown by using a subset of randomly selected independent variables. As we are dealing with a regression problem, the final result is computed through the mean of the outputs provided by the decision trees (Gong et al. 2018; Loureiro, Miguéis, and da Silva 2018). The algorithm uses an ensemble meta-algorithm in order to create the random choice of trees, named as bagging (also known as bootstrap aggregating) (Tech 2016). This kind of selection is what differentiates the Random Forest's robustness from the decision trees' one. These models have higher accuracy and have a good capacity of dealing with big datasets, as they are not as sensible to little variations (Loureiro, Miguéis, and da Silva 2018).

2.3.2 Gradient Boosting

Gradient Boosting is an algorithm specially characterized by the way it learns. It makes use of a sequential model construction that builds new models which learn from the shortcomings of the previous ones. To define its configuration four inputs are required from the analyst. One of them being training data, a vector containing the input variables and the target variable. The second one would be the stopping criteria, which includes the number of iterations or a minimal error improvement. It is also required to choose a loss function and a base-learner.

Regarding loss functions, there are several available for selection and they are constrained by the data type of the output variable. In case of continuous target variables, squared or absolute error are usually selected functions. The loss function aims at measuring the model predictive performance error in order to minimize it. When it comes to base-learner models, these can be linear or smooth models, decision trees or other functions. Frequently decision trees are the most adopted. Figure 4 shows the sequential learning process associated with trees creation (Natekin and Knoll 2013).

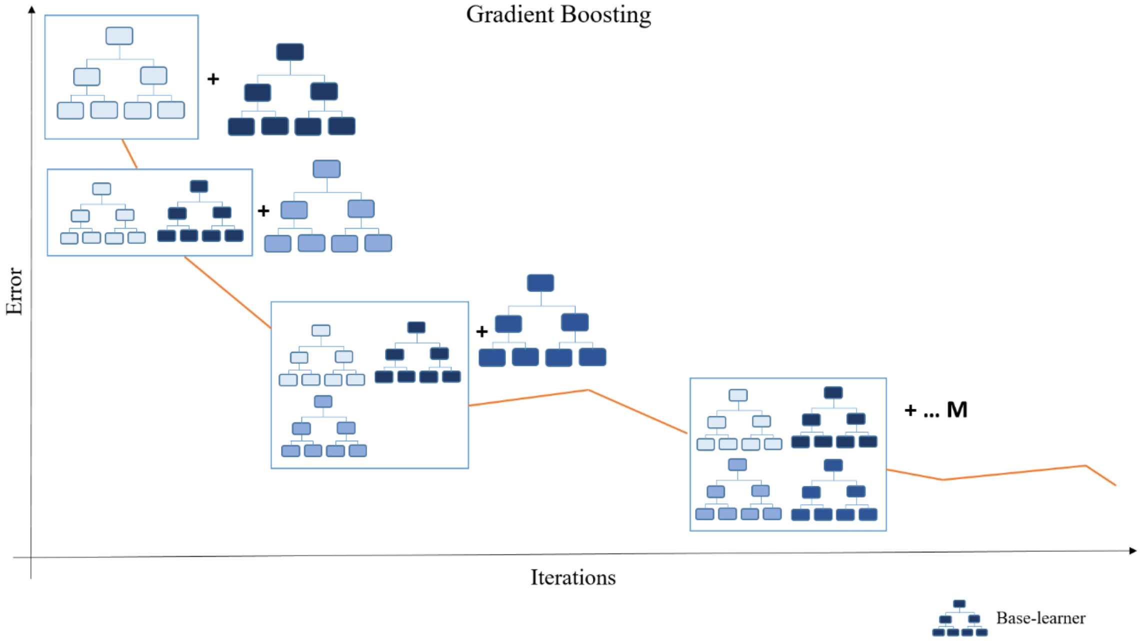
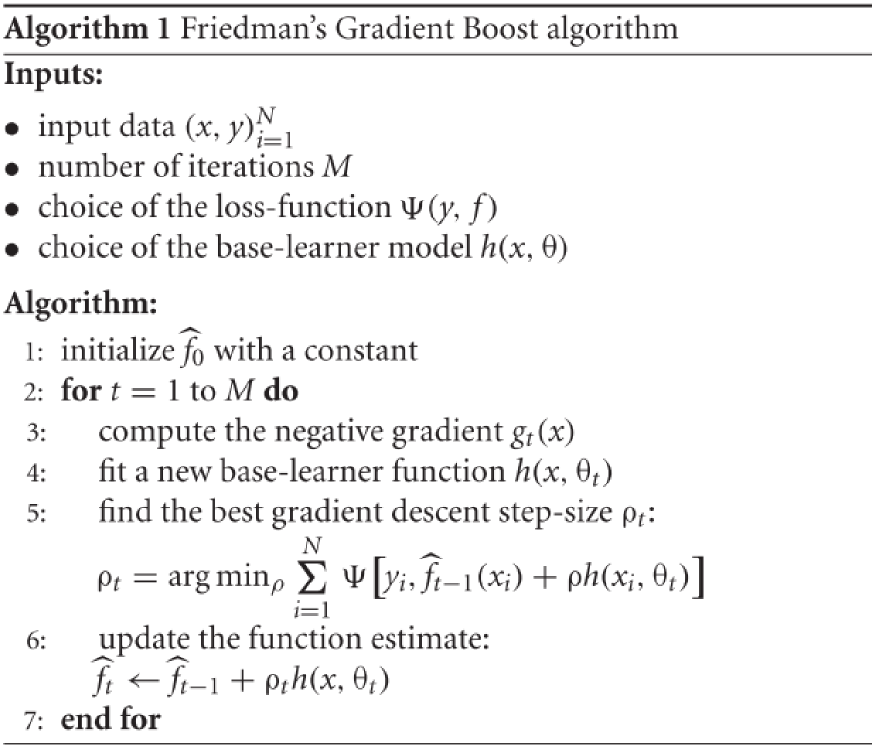


Figure 4 – Gradient boosting learning progress representation based on (KDnuggets 2019)

In order to reach the best possible prediction values, the loss function is minimized using a gradient descent optimization. Gradient descent defines the direction in which the function decreases at the highest rate. Thus, the negative gradient of the loss function is computed so it can achieve its local minima in the fastest way. This approach makes use of a stepwise procedure, where the step-size can be defined differently for every leaf of the trees. Figure 5 shows pseudo-code describing Friedman's Gradient Boosting algorithm



Algorithm 1 Friedman's Gradient Boost algorithm

Inputs:

- input data $(x, y)_{i=1}^{N}$
- number of iterations M
- choice of the loss-function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm:

```
initialize $\widehat{f_0}$ with a constant
for $t = 1$ to $M$ do
    compute the negative gradient $g_t(x)$
    fit a new base-learner function $h(x, \theta_t)$
    find the best gradient descent step-size $\rho_t$:
    $\rho_t = \arg\min_\rho \sum_{i=1}^{N} \Psi\left[y_i, \widehat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\right]$
    update the function estimate:
    $\widehat{f_t} \leftarrow \widehat{f}_{t-1} + \rho_t h(x, \theta_t)$
end for
```

Figure 5 – Friedman's Gradient Boost algorithm in (Natekin and Knoll 2013)

According to Figure 5, the user starts by defining the number of iterations, the loss function and the base-learner model. Then, he assigns a constant value for the initial estimation, usually the mean or median of the observed values. Then, begins an iterative process where the loss function gradient is computed and the new base-learner function fits the data. The gradient descent step-size is then calculated and finally the initial estimate is updated by the new one. This process is repeated for M iterations (Natekin and Knoll 2013).

2.3.3 Artificial Neural Networks

Artificial neural networks are currently one of the most explored data mining techniques due to their good accuracy and capacity of identifying and predicting uncommon patterns. Despite that, the long running time and little intuitive interpretability are some of their shortcomings. This algorithm is called neural networks, since it is composed by a set of units that serve as input and/or output and are connected to each other. The typical structure of an ANN is composed by 3 or more layers, the input, output, and hidden layers in the middle. The number of layers, along with the number of neurons define the network's topology, and is chosen by the analyst, usually through an iterative process. It can become a very complex and time-consuming process to choose the "right" values for these parameters. Figure 6 shows the structure of ANN. The first layer receives inputs, and they are transferred to the following layer until the last one provides the final outputs (Han and Kamber 2001). The review of ANN was based on *Data Mining: Concepts and Techniques* from Han and Kamber (2001).

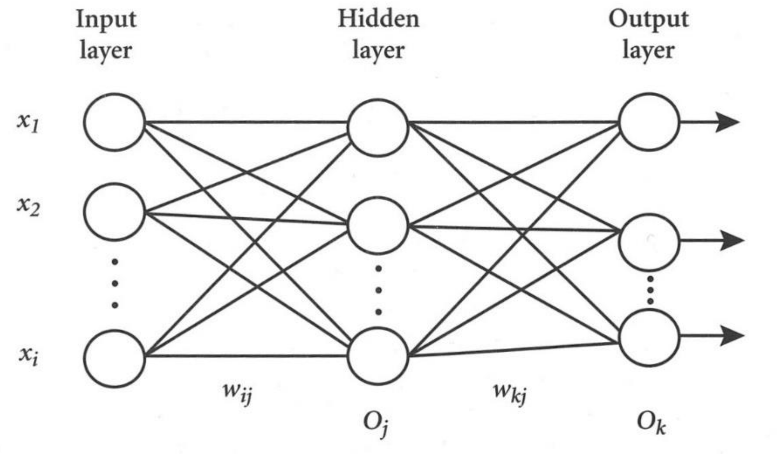


Figure 6 – Neural Network structure in (Han and Kamber 2001)

A set of weighted connections relates every neuron to all the ones present in the next layer. A bias is also associated to every neuron and it considers all connections from the previous layer, defining a threshold for every connection. For creating this model, training data is provided to the network, so it can adjust its weights and biases to best represent the data. Usually, this kind of learning technique benefits from receiving big amounts of data. However, this may lead to overfitting, and instead of representing the population, the model represents only the sample's behavior. As the outputs only go to the following layers, this is considered a feedforward mechanism.

The learning process, however, may be performed according to different algorithms. The most commonly used is backpropagation and it "learns by iteratively processing a set of training samples, comparing the network's prediction for each sample with the actual known class label"(Han and Kamber 2001, 305). For each training sample, the weights are modified to minimize the mean square error (Equation (2.5)) between the network's prediction and the actual class. The computation of this error goes from the last layer until the first (backwards). The weights are expected to converge at a given moment. Equation (2.4) represents the input for unit j as the weighted sum of the i output connections to the previous layer plus the bias, commonly known as the activation function. This is schematically represented in Figure 7.

$$I_j = \sum_i w_{ij} Oi + \theta_j \qquad (2.4)$$

Where:

I_j, is the input for unit j
w_{ij}, is the weight of connection with unit i from previous layer
O_i, is the output of unit i
θ_j, is the bias

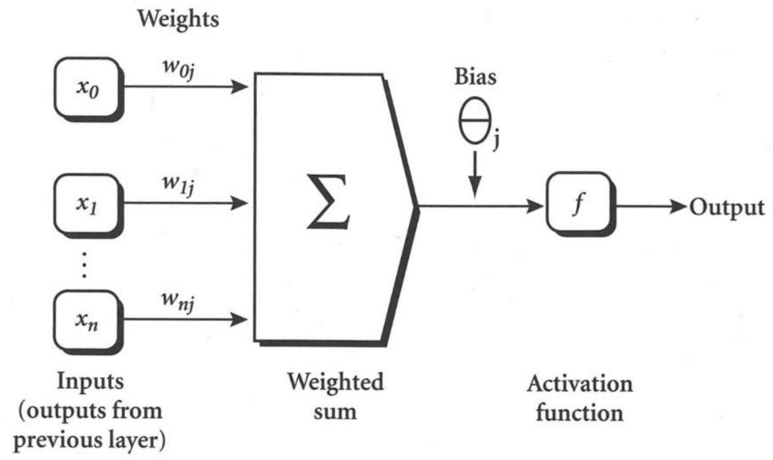


Figure 7 – Activation function scheme in (Han and Kamber 2001)

$$Err = O_j(1-O_j)(T_j-O_j) \tag{2.5}$$

Where:

Err, is the Error
Tj, is the true output value A
Oj, is the output of unit j

As the error is propagated, also the weights and biases are updated, using a learning rate that considers the error. The learning rate is a value defined between 0 and 1 that helps finding the global minimum, in smaller or larger steps according to its value. Their update functions are shown in Equations (2.6) and (2.7).

$$w_{ij}=w_{ij}+l*Err*O_i \tag{2.6}$$

Where:

w_{ij}, is the weight of the connection between neurons i and j
l, is the learning rate
Err, is the Error
Oi, is the output of unit i

$$\theta_j=\theta_j+(l)Err_j \tag{2.7}$$

Where:

θ_j, is the bias of unit j
Err, is the Error
l, is the learning rate

These equations ((2.6) and (2.7)) are applied after the presentation of each observation; thus, this is called case updating. On the other hand, the analyst may decide to update only after all samples are used and accumulate these values in separate variables. That technique is called epoch updating.

Finally, the algorithm stops when it gets bellow the defined threshold for change in weights, (the second member of Equation (2.7)), or the frequency of wrong outputs obtained in the epoch is lower than a predefined threshold or a given number of epochs was completed (Han and Kamber 2001).

2.3.4 Performance evaluation metrics for regression

After applying the previously mentioned algorithms or any other machine learning model, a performance evaluation needs to be done, to measure how good the model predicts. For regression problems, some of the error measures commonly used can be seen in Equations (2.8), (2.9), (2.10) and (2.11). Regarding these measures, they all aim to compare the predicted values with the real ones, commonly mentioned as observed values. In the case of MAE and MAPE, they work with the absolute difference. The other two measures square it. This difference is called the residual difference, since it takes the residual error from each observation and its estimated value. Regarding MSE and RMSE, these are more sensitive to outliers, since they assign weights to the residuals. A good model aims to reach the minimum values of these measures as it indicates their predictions are closer to the real values. (Swalin n.d.)

The awareness about the order of magnitude and range of target variable is crucial for results comprehension, since, a given measure can be considered high if its baseline is much smaller. However the same absolute value can represent a much smaller error when the order of magnitude of its baseline takes bigger values.

The presented measures do not have an absolute scale such as R-squared for example, as they are measured in the units of the target variable. Thus, a better model can be chosen by comparing it to other models or to the observed values range and average for instance.

$$\mathrm{MAE}= \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left|\mathrm{y_i}-\hat{\mathrm{y}}_{\mathrm{i}}\right| \tag{2.8}$$

Where:

MAE, is the mean absolute error
yi, is the observed value
$\hat{y}_i$ is the predicted value
N, is the number of observations

$$\mathrm{MAPE}= \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left|\frac{\mathrm{y_i}-\hat{\mathrm{y}}_{\mathrm{i}}}{\mathrm{y_i}}\right| \tag{2.9}$$

Where:

MAPE, is the mean absolute percentage error
yi, is the observed value
$\hat{y}_i$ is the predicted value
N, is the number of observations

$$\mathrm{MSE}= \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{N}(\mathrm{y_i}-\hat{\mathrm{y}}_{\mathrm{i}})^2 \tag{2.10}$$

Where:

MSE, is the mean squared error
yi, is the observed value
$\hat{y}_i$ is the predicted value
N, is the number of observations

$$\mathrm{RMSE}= \sqrt{\frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{y_i}-\hat{\mathrm{y}}_{\mathrm{i}})^2} \tag{2.11}$$

Where:

RMSE, is the root mean squared error
yi, is the observed value
$\hat{y}_i$ is the predicted value
N, is the number of observations

3 AS-IS Situation – Context and company's solution

This chapter aims to explain the assortment planning process and provide a concrete example from a widely used application for this purpose, framing the item attribute weight challenge using a solution of Oracle Retail and showing what is currently made. Other software solutions could have been used, since assortment planning is transversal to all fashion retailers.

3.1 Assortment planning

According to Oracle Retail user guide, "Assortment Planning is a business function that merchants and planners perform to determine the appropriate mix of products that maximizes organizational goals: sales, profits, inventory turn, and so on" (Goodman 2016, 1-7).

Several planning tasks are required for a retailer to launch a new collection. Activities require supervising along the full supply chain, making it important to have an organized platform where managers can decide, communicate and save information regarding all processes from the moment an item is ordered to the supplier until it is sold at the final customer. The definition of profit goals, assortment selection, allocation of product to specific channels, getting familiar with fashion latest trends and many other activities require time, dedication and supervising, since increased efficiency can be gained through the use of a good decision support system tool. It is in this context that assortment planning is framed, being one of the many required steps to increase the companies likelihood to be successful.

The example of Oracle Retail is explored in higher detail, afterwards, in order to gain a more practical sense of how these systems are used, organized and applied nowadays.

3.2 Merchandise hierarchy

Typically, during the planning process, decisions about the products need to be taken at different levels of aggregation. For instance, in an early stage, the planner wants to define that women's collection from the Spring/Summer season is going to be composed by an X number of T-shirts and a Y number of trousers. In a later moment it becomes necessary to define which models and colors will be chosen from the trousers category to have in store Z. At the next phase, one must decide which sizes of those specific trousers will be present on that store. This phenomenon leads to a common practice of creating a hierarchy to distinguish the level of aggregation one is dealing with. Several terminologies can be used. Figure 8 shows an example of designations that may be adopted for this purpose. Some of this nomenclature is present on this project in the datasets provided for analysis.

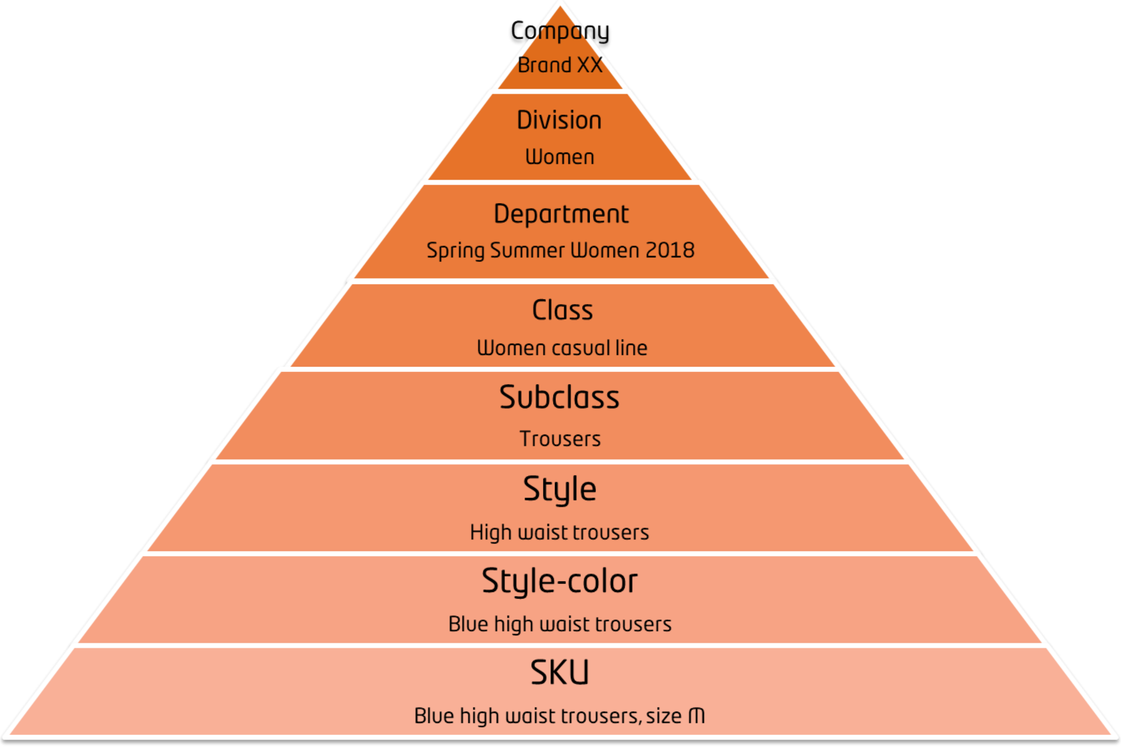


Figure 8 – Merchandise hierarchy

Regarding the concepts present in the analyzed dataset, it is important to be familiar with the definition of SKU, stylecolor, look group and look. A SKU – Stock keeping unit – represents the lowest aggregation level at which a product can be described, containing all the information describing the product type, characteristics, color, size and sometimes the season. Stylecolor is the level that comes right after SKU, identifying the second lowest product aggregation level, with the only difference of not discriminating the size. When it comes to the concept of look group, it usually refers to a campaign period in calendar, for instance to have the Spring/Summer collection from February to August. Looks are a way of making shorter time divisions contained into look groups, identifying not only the timeframe, but also the products available for sale on that time. A look group is composed by several looks. These concepts were created to make it simpler to manage the season.

3.3 Oracle Retail tools for assortment planning

3.3.1 Integrated fashion planning and optimization overview

Oracle Retail’s integrated fashion planning and optimization software is composed by several smaller modules, including the Assortment Planning one. These modules can be seen on Figure 9. The first one aims to create a Merchandise Financial Plan, where strategies are defined focusing mainly on the capital perspective of the business. Afterwards, retailers must plan the assortment of the product offering, along with all managerial decisions it implies. At Item Planning, there is a lower decision-making level followed by the definition of all promotional events that will happen in the time period considered. Clearance and Optimization are next accomplished, followed by Size Profile Optimization (Goodman 2016).

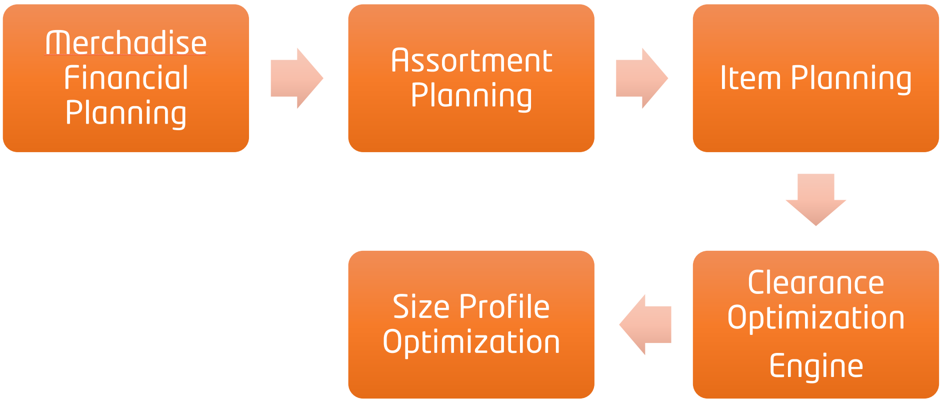

Merchadise Financial Planning
Assortment Planning
Item Planning
Clearance Optimization Engine
Size Profile Optimization

Figure 9 - Integrated fashion planning and optimization in (Goodman 2016)

This project's motivation is to fill a gap regarding the assortment planning process. For that reason, a better understanding of this concept and associated software is necessary.

3.3.2 Assortment planning step by step

Assortment Planning is a module directed to several users with different management responsibilities. These people can be divided into four groups: the buyers (senior and regular buyers), store informants (including the visual merchant and the sales planner), merchandise and allocation. Table 1 shows all steps included on the assortment planning business process along with the responsible assigned to them.

Table 1- Assortment Planning business process tasks and responsible in (Goodman 2016)

Process Step	Senior Buyer	Buyer	Visual Merch	Sales Planner	Merch Planner	Allocator
Plan Setup		Aligns looks with Calendar	Sets up Fixtures and Capacities	Supports Planning Process	Creates the Merch and Local Plan	
Selling Curve Maint				Maint the curves		
Cluster Maint		Builds/Updates Cluster				
Create the Shopping List	Determine Strategy & Fills					
Fill the Wedge		Sets Wedges Strategy and Assort				
Buy Planning		Creates and Re-trends the BuyPlan			Review Buy Plan	
Size/Pack Allocation		Makes Size/Pack Decisions				Uses Allocation by Size/Pack
In-Season	Review in-season trend and plans				Review Sell-Thru and BuyPlan Changes	Executes Inventory Strategy

- Plan Setup

In this phase, all program's configurations must be done, so it can be prepared to receive the information users will provide as an input. The information is mostly related to store attributes and their values, targets to achieve, packs, available capacities and so on.

Looks and look groups are defined in this phase too. For planning the setup, strategies and item sets are assigned to looks and look groups. Besides that, goals are set for each look, through definition of the percentage of styles that are part of the assortment by attribute and color. For example, a look can be defined by having denim products, composed by 50% of shirts and 50% of jeans. Supposing this look group is assigned to a cluster of stores with 5 stores, where each will receive 100 items of this look. This would mean that each store would receive 50 denim shirts and 50 jeans to have exposed in the store for that period of time. The decision regarding how many items to assign this look and the assortment that composes it, will be reflected on the product the client finds on the shelves. So, if it is correct, the selected composition of products will correspond to the demand, having no over or understock. This would be the ideal for any company to achieve as it would represent a maximization of sales and no sunk costs. However, if it is wrong, sales opportunities can be lost and products can become overstock as they will not satisfy customer demand.

This is a very important step, as these percentages correspond to the weights of attributes that are being calculated on this project. Currently, these weights are chosen by the user, based on his experience, his perspective of the market and the conclusions he takes from going through data of items with similar behavior from past seasons, for instance. The objective of this dissertation relates to finding those weights based on data-driven predictive techniques.

- Selling Maintenance Curve

In this step, the sales planner can set manually or load from external systems, product lifecycle curve for new item selling patterns, according to similar products that existed in the past.

- Cluster Maintenance

Clusters of stores are defined in this step, according, either to the store performance (related to the strategy chosen in the looks phase, ex: sales growth), and/or according to store attributes, such as store size or temperature.

- Create the Shopping List

The Shopping list is where the senior buyer makes strategic decisions, working at a highly aggregated level of assortment. First, historical data is analyzed. Afterwards, the assortment size (number of styles) is set and the shopping list's goals are defined, without disregarding the need to guarantee the financial objectives and to respect the constraints defined before. It's also important to maintain the attributes and customer segment goals previously defined when planning the setup. However, they can be changed upon reconsideration. Finally, the list is created and then approved. During this process, relevant decisions are taken, including the definition of number of colors each style will carry and its sales potential, the quantities to buy from each item category and the selling goals to achieve.

- Fill/Build the Wedge

A wedge is a concept frequently used in retail due to the picture created when showing the dimension of the assortment grouped by clusters creating the shape of a wedge. This step is meant to perform the assortment with an increased detail, so it refines the Shopping List created before. Adjustments are made to the previously defined strategies, in order to accommodate the lower level conditions. Besides that, the stylecolors are chosen here.

- Buy Planning

  The buying plan defines the projection of sales over time and the receipt and inventory plan. It considers the initially defined targets, along with the assortment decisions made during the process and tries to allocate timings and quantities and predict sales and costs associated to different PoC. The receipt and inventory plan includes inputs from Build the Wedge step to which lead and lag times are assigned.

- Size/Pack Allocation

  This is where optimization of product orders is made, taking into account packaging and product dimension restrictions.

- In-Season

  At this point, the season has already begun and product is already spread around different PoC. Here, the performance of every item, store, look, and so on is reviewed and evaluated. If considered necessary, changes are made. (Goodman 2016)

3.4 Problem definition

The item attribute weights in this specific context are defined based on the managers' experience, who provide a manual input for each of the weights. The object of this project is to provide a data-based suggestion for fashion companies to use and maybe tools such as the one from Oracle Retail may integrate in their solutions. The method proposed to explore this question, including visual analysis, models creation and feature importance prioritization can contribute to the purpose of better predicting fashion sales and understanding if the items physical characteristics have a high impact on them.

4 Proposed solution

This chapter aims to describe the methodology adopted to reach the proposed solution. Data from two years from a Mexican multibrand fashion company, named here Company X, was used. This company works with several brands simultaneously and has clothes, accessories, shoes and *lingerie* for all ages. It also has stores spread all around the country. A better description of the company's available dataset is given in subchapter 4.1.

The methodology proposed involves three main steps. The first one is to understand, clean, organize and filter the data making use of the MySQL infrastructure to handle the database. The second step involves a visual data analysis making use of Tableau software to look for correlations of predictive variables. The third phase involves testing the performance of three machine learning models using Random Forest, Artificial Neural Networks and Catboost algorithms to identify relevant predictive variables according to their influence in sales. For that, RStudio and Spyder were used, exploring different packages that better suited the purpose. Finally, a proposal of criteria selection is made through a combination of the last two steps.

4.1 Data preprocessing

4.1.1 Variables description

Table 2 shows the final data obtained after the preprocessing phase, its type and a brief description. It also identifies the class which variables belong to, for better comprehension of the dataset. All predictive variables belong to the categorical type. However, three different scenarios regarding the target variable definition were tested on the models. Therefore, the table presents three output variables, in the Final Sales class, all of them numerical. Further explanation on how they were computed is given in section 4.1.7.

Table 2 – Variables' description

Class	Name	Type	Description
Time	Year	Ordinal	Year when the sale happened
Product attribute	Brand	Categorical	Identifies the brand the item belongs to
	Color	Categorical	Identifies the color of the product
	Fashionability	Categorical	Describes the product according to how it follows recent trends
	Gender	Categorical	Describes the target consumer gender
	Price label	Categorical	Identifies the price range of the product
	Product family	Categorical	Identifies the description of the product's use
	Product type	Categorical	Groups product family according to product function
Sales condition	Sale type	Categorical	Distinguishes the sale context on which the product was sold
Final Sales	Sales Units	Numerical	Total units of product sold in Year X, under Sale type Y
	Sales Units per Week	Numerical	Total units of product sold in Year X, under sale type Y per week
	Ratio	Numerical	Percentage of time with sales + Percentage of sales over average + Percentage of full price sales

4.1.2 Raw data

Data was extracted from the records saved on Oracle Retail's application and was registered into four main fields: hierarchies, attributes, sales and assortment.

Hierarchy data describes all possible aggregation levels with respect to calendar, location and product. This kind of table contains information from the most detailed level of each of these topics - day, store and stylecolor, respectively, until the most generalized. Their structure can be seen bellow.

- CALENDAR: **Day** | Week | Week of the year | Month | Quarter | Semester | Year | Holiday | Event
- PRODUCT: **Stylecolor** | Style | Subclass | Class | Family | Department | Group | Company | Clas | Gtot | Tsub | Itg1 | Itg2 | Itg3

- LOCATION: **Store** | Ccpt | Region | Country | Channel | Corp | Tstore | Tcon | Bstore | Oreg

Attributes information was provided on product and store with the following variables, respectively:

- PRODUCT**: Stylecolor** | Color | Company cost | Initial Price | Final Price | Consumer type
- STORE: **Store** | Channel | Region | Store Type | Weather

These store attributes were disregarded due to their incoherence with the location hierarchy.

Sales information was registered at two different product aggregation levels: SKU and stylecolor. The information regarding stylecolor sales was recorded distinguishing three different kinds of sales - Regular, Promotional and Clearance. Regular sales can be described as when a product is sold at its standard price, without any sort of campaign or change in price appealing for it to be bought. Promotional sales, on the other hand, are the exact opposite. They refer to products that have been subject to strategies such as changes in price, advertising or special location assigned in the stores, to catch customer's attention and foster sales. Clearance sales happen when the product is identified as a sunken cost to the company, by not having sales, and thus becoming unsold stock. Other special situations may lead to selling products in clearance. In these occasions, its price is lowered until the product is sold, since the company no longer expects to make a good profit with those products.

The data provided at SKU sales level did not record the different types of sales, so the decision was taken to consider the stylecolor as the lowest aggregation level for the product.

When registering sales, it is important to refer the conditions in which the sale was made, regarding not only the product being sold but also where, when, at what price and in what sale circumstance. For that reason, the sales table is structured as follows:

- SALES: **Stylecolor | Week | Store** | Regular Units | RegularRetail | Promotional Units | Promotional Retail | Clearance Units | Clearance Retail | Cost Retail

Six of the variables from this table have the first name referring to the sale type performed (Regular, Promotional or Clearance) and the second identifying if it refers to the number of units sold – Units - or the value of sales in monetary units (Mexican pesos in this case) - Retail. Each observation was recorded by store, week and stylecolor. Zero-valued sales were not registered. Cost Retail refers to the cost of the product for the Company X.

The assortment information aims to inform about the stylecolors available for sale in a specific group of stores and period of time. This information could not be used as it didn't have data for all stylecolors variety present within sales records.

- ASSORTMENT**: Look** | Look group | **Store cluster** | **Stylecolor**

Another problem in the assortment data relates to the fact that, most often these looks do not correspond to the real time products are in store. Even though there is an initial assortment plan, the most likely situation to happen is for the product to be longer or shorter periods in the stores as a consequence of manager decisions upon verifying the sales behavior in real time. This way, many changes to these initially defined assortments are made. Thus, they don't describe the true behavior of an article in a store.

An important detail regarding the assortment refers to the big variety of existing look groups due to the different seasons considered for each of the brands of Company X, the target consumer and the store cluster. For instance, a possible look group can be "Autumn/Winter Man Brand 1 Cluster A". So, a very high number of possible combinations could be taken from here.

4.1.3 Lack of information

Despite the high number of variables present on the raw data, many of them were eliminated from the dataset due to the lack of filled-in data. For that motive a rule was created defining a minimum threshold of 70% of non-NA data for a variable to be considered for the final dataset. This led to a big reduction of data.

Another criterion that defined the lack of information of an input is its relevance. A variable was considered relevant if it would add value to describing the product's characteristics and conditions in which sales were performed. Keeping in mind the final goal of understanding which product criteria could be used for a suggestion of assortment that will help maximize sales is, therefore, the core filter to choose variables to keep. For example, color is a variable that contributes to the product description, so it was kept. However, company cost is not an information that has a direct impact on sales or consumer's interest in buying a certain item. Thus, it was eliminated.

4.1.4 Data aggregation

This project aims to implement a solution to support high level decisions on assortment planning. Thus, it becomes necessary to define in which timeframe, space division, product level and target consumer to decide.

Starting by the timeframe, the first option, which is more intuitive, is to predict these criteria for season (Spring/Summer, Autumn/Winter). Firstly, this option was considered. However, after looking at data it was possible to see that this seasonal division does not make sense for Company X, due to the timespans products remain in stores and the high variety of weather conditions found in different regions of Mexico. When it comes to the product, data shows that some products cross seasons and for that reason, this separation would probably not make sense. In this dataset, product attribute fashionability comes from the Subclass classification of items originally in the raw data and the sort of categories it considers is commonly used to describe the lifecycle of a product, where usually a basic item stays for longer periods in stores while fashion items have short life cycles. Figure 10 shows that despite products having different classifications for fashionability, most of them sell for a small number of weeks, and the remaining products can go up to 52 weeks from the first to the last registered sale. So, fashionability does not relate only to product lifecycle.

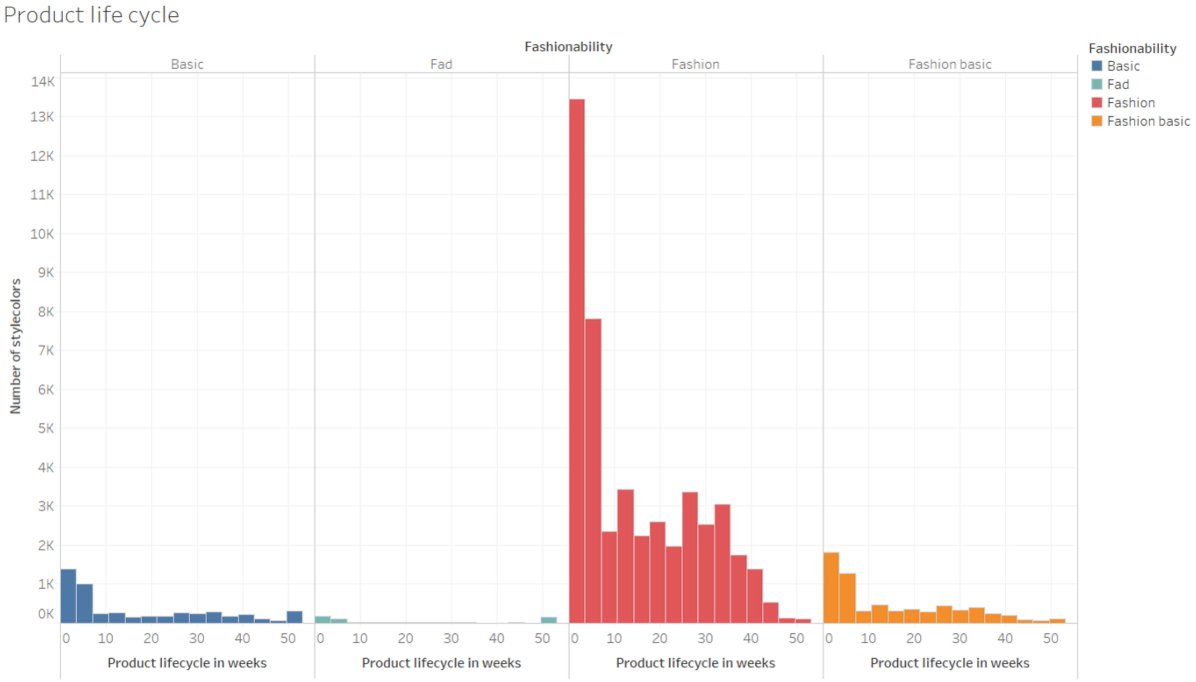


Figure 10 – Product life cycle histogram

The second reason for the time aggregation level choice relates to the fact that stores are spread all around Mexico which means that the weather from one store to the other can vary a lot, since temperatures and rain conditions can be very different from region to region (Barbezat 2018). Therefore, the decision was to predict for a whole year. Company X defines a year starting in February and going to February of the following year. This means including 2 seasons that complete a year. So, a year is the time aggregation level from February to February defined for this project.

Concerning geographic space, sales could be evaluated by store, region or all together. Considering the kind of decision being made, stores were considered all together since the assortment planning module has a later store assignment and clustering section that takes care of the store allocation.

Regarding the product and its target consumer, the company has a classification that distinguishes brands, from A to E, considering three variations of brand A, which represents almost half of the total sales. Company X also considers seven target consumer groups including Women, Man, Babies and Kids. Within these groups, there still is a classification into Accessories, Shoes and Underwear which despite being classified within the target consumer groups, include all types of consumer, differing only on the kind of product being sold. In order to choose a more homogeneous target consumer and products with more similar behavior, only clothes were considered for analysis, so accessories, underwear and shoes were discarded. Sales for different target consumers are presented on the graph from Figure 11, distinguishing brands. The values are merely representative of the real sales.

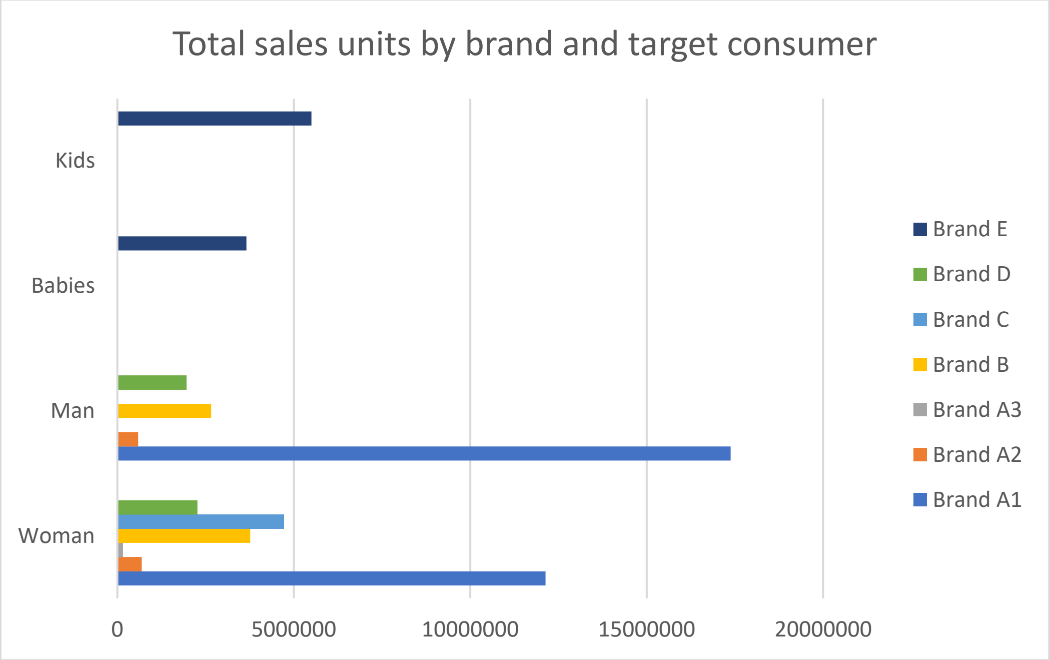


Figure 11 – Total sales units by brand and target consumer

Looking at bar plot on Figure 11, it is possible to see Brand E is the only having sales for kids and babies, presenting lower values of total sales, so it was decided to analyze sales for man and women of all brands.

Another classification is made regarding the product which relates to its family and type. These classifications are done with respect to two different product aggregation levels, where product type is a higher aggregation level such as clothes to wear above the waist, named as tops for example, which include several product types such as T-shirts, jackets, shirts and so on. The choice regarding the product aggregation was to go through an iterative process in which first all products are evaluated together, and afterwards evaluation is refined by product type or even by analyzing individual product families with higher sales.

4.1.5 Outliers analysis

The provided dataset contained three types of numerical variables, namely units sold - Sales Units (Regular, Promotional and Clearance), sales value - Sales Retail and Price. Despite the price information being provided in the product attribute's table, through variables Initial price and Final price, these values corresponded to a value of the product price recorded in the moment of its first sale and at the last time it was registered, respectively. However, changes in price may occur during time, leading to different kinds of sales. For that motive, price was instead computed through the quotient of Sales Retail records by the respective units sold. As these values were recorded along two years, it is possible to look for abnormal behavior by analyzing how they perform through time. This is a common practice in preprocessing, so peaks and low values, cyclic behavior or trends can be identified by the analyst, understanding if they had a cause, if they were one-time events or even if they are just a consequence of a human mistake.

Looking at the scatterplots bellow in Figures 12, 13 and 14, it is possible to see that in in Regular sales there is a similar shopping behavior across gender. In Regular sales, there were peaks of sales which happened repeatedly in the two consecutive years and can be explained by Christmas time and *El Buen Fin*, the Mexican black Friday in November.

When it comes to promotional sales, Mother's Day marks a very strong value of sales in April/May for women. In the other hand, for man, June is a month that shows high promotional sales repeatedly. By studying the brand's behavior, it was possible to understand the remaining promotions happen without a clear pattern or schedule. It seems promotions on some items are decided at given times of the year when Company X decides that is good. This represents a possible cause to some of the observed peaks.

Clearance sales show some seasonality, despite the very unstable data. In Company X there are products being sold in clearance during the whole year, which can explain the relatively constant pattern. These sales are, however, very unpredictable.

It is possible to see a big difference in the order of magnitude of the sales registered between sale types. Regular sales are clearly the prominent ones, followed by clearance and finally promotional.

When it comes to Sales Retail, this variable was not used for prediction, so the outlier analysis was not performed on it. Price on the other hand, was analyzed, and most values corresponded to the threshold of prices usually put into practice in Company X's brands. While doing this comparison, one observation recorded an extremely high price. Considering the other sales registered for the same product at the same time in other stores, when different prices were found, the decision was made to change this item's outlying price to the same price the product had on other stores, assuming there was an error when registering Sales Retail with an oversized value. All the other values were maintained.

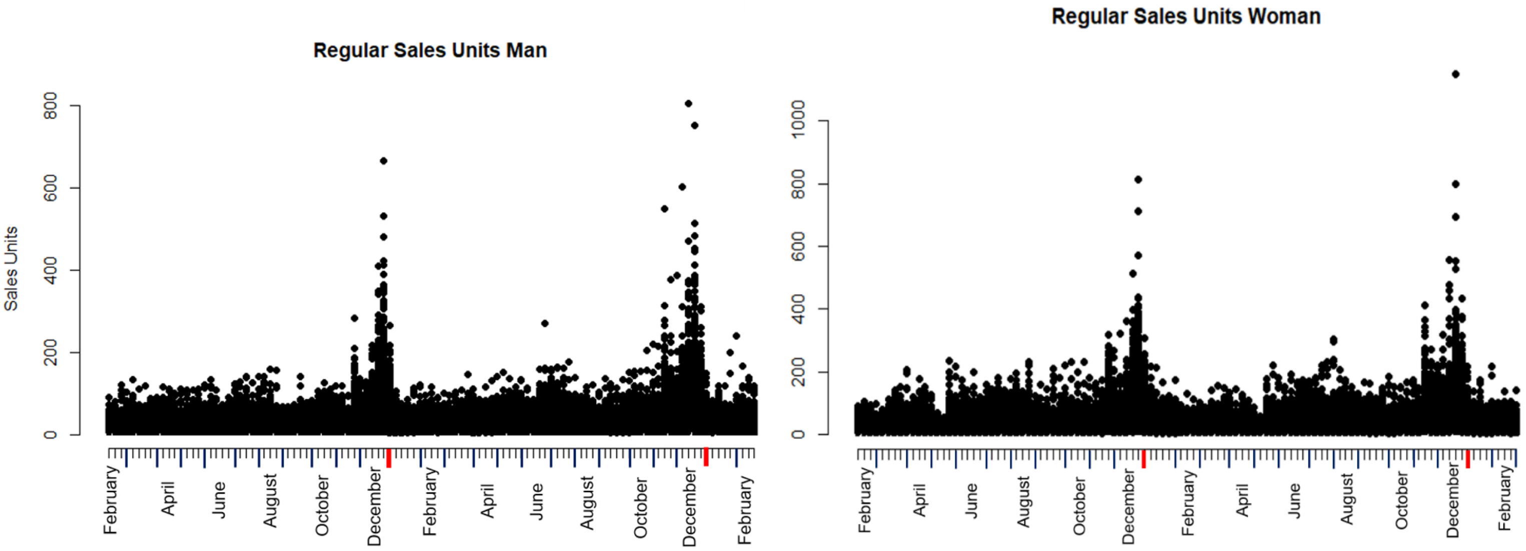


Figure 12 – Regular sales units across gender

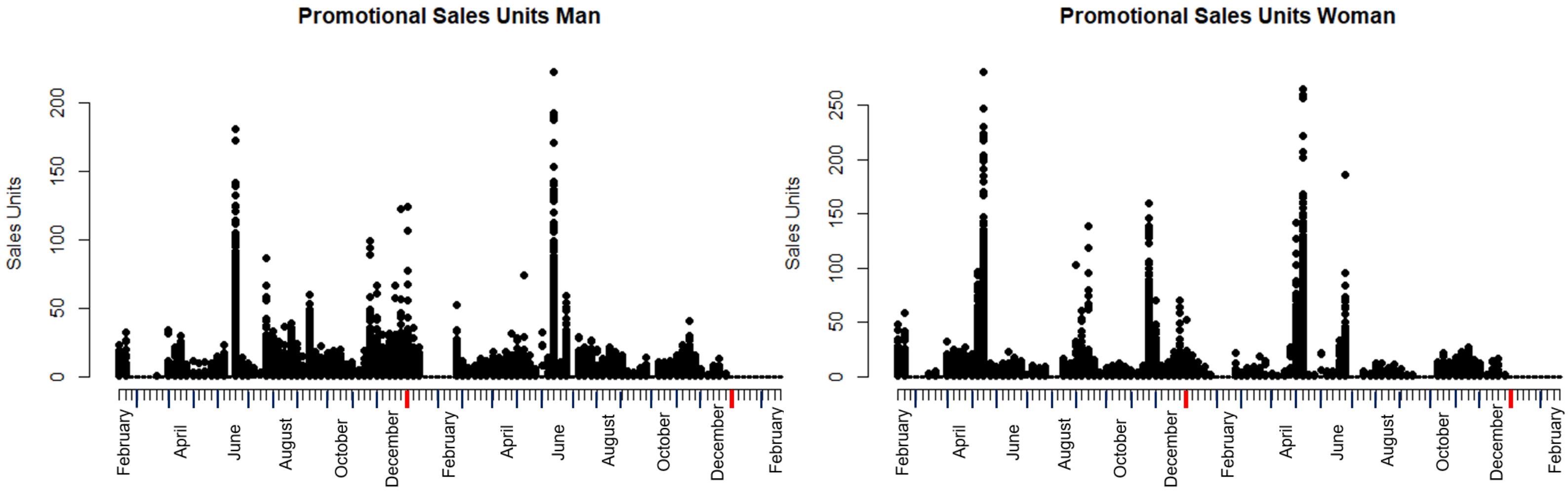


Figure 13 – Promotional sales units across gender

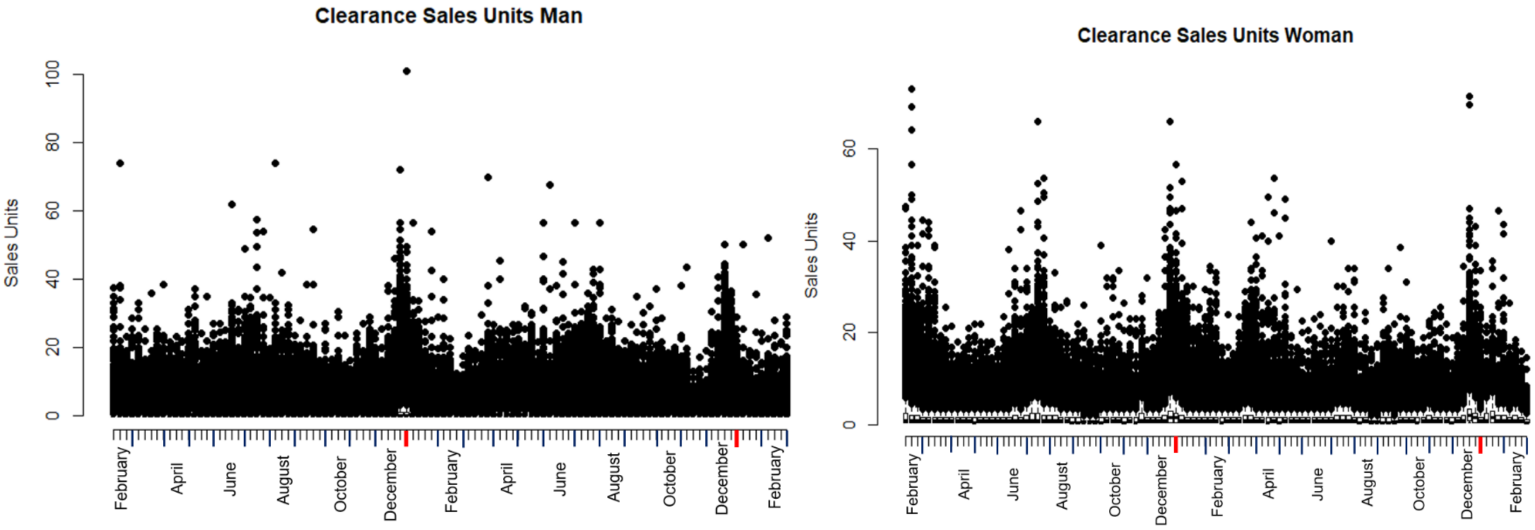


Figure 14 – Clearance sales units across gender

4.1.6 Variables categorization

Most of the product attributes can be described through categorical values. However, a variable with way too many categories does not provide good information for a predictive model to learn. Thus, it was necessary to group the variable Color as it presented a way too extensive list of possible tonalities.

Price was also categorized. After computing price, it was possible to see many different values varying very little from each other. For a better understanding of its impact, a decision was taken to create five different price ranges. The intervals' range was defined by the Equation (4.1).

$$PI=\frac{Max(P)-Min(P)}{5} \tag{4.1}$$

Where:

PI, is the price intervals range
Max(P), is the highest price of all products
Min(P), is the lowest price of all products

Ranges were classified in five levels: "Low", "Medium low", "Medium", "Medium high" and "High" and they were computed considering prices from all products in analysis.

4.1.7 Output variable

The methodology chosen for this project involved a visual analysis and predictive model creation from the sales data. When it comes to the modeling phase it is important to make several iterations to understand which best fits the problem to be solved. The goal was to understand how predictive variables – item attributes - impact sales and for that purpose three different scenarios were constructed in order to compare their results. These are considered to be possible scenarios in a sense that they represent the potential choice of output variable to use of the analyst.

The first scenario simply predicts Sales Units by year, in the whole country, for a specific type of sale (regular, promotional or clearance).

The second scenario tries to consider sales over time, where sales units are divided by the total number of weeks the item was sold in that type of sale. Ideally, the time products were on sale in the stores should be taken from the assortment data. However, very little stylecolors were recorded on that dataset. So, the assumption to calculate that time was to consider first and last sales of the product in the given year. For example, if a T-shirt is sold in a regular sale for the first time in the 13th week of 2017 and registered its last regular sale in the week 33 of 2017, then, the regular sales units per week of that t-shirt will be the total units sold over 20 weeks. This scenario was built in order to consider the time products were on sale in a given sale type. Sale type is a very important variable since it reflects how fast product flows out of the stores. It means that if a T-shirt is sold only on regular sales, then it is a product that sparked the client's interest. On the other hand, if it was sold in clearance, this would mean that it had not been successful by itself and thus, was likely sold only due to its low price. The lack of information about store stock can also lead to an increasing error on the predictions of this variable.

For a third scenario, a ratio was considered as the output variable. This ratio aimed to account for the interest of clients on a specific product and how good it was, by itself. The ratio considers three measures: sales, time and full price and is computed for a stylecolor in a certain year, as shown in Equation (4.2).

$$\mathrm{Ratio}_i = \frac{SU_i}{\frac{1}{N}\sum_i^N SU_i} + \frac{\mathrm{Count}(Wi)}{\mathrm{Max}(W_i)\text{-}\mathrm{Min}(W_i)} + \frac{RSU_i}{SU_i} \qquad (4.2)$$

Where:

SU_i, is the units sold of stylecolor i
N, is the total number of products sold
Wi, is the index of the weeks when product I was sold
RSU_i, is the regular sales of stylecolor i
Count(Wi), accounts for the number of weeks stylecolor i was sold
Max(Wi)-Min(Wi), accounts for the total number of weeks stylecolor I was exposed for sale

A higher ratio represents higher customer's interest on the stylecolor which makes it a desirable item to repeat in the future. For this model, conclusions can be taken about the best criteria for maximizing sales, directly from the products with the highest ratio. The first term of Equation (4.2) represents a margin of product sales over the average sales of all products in that year. The second term refers to the number of weeks the product registered sales over the total time it was exposed for sale. The denominator total time was computed through the assumption that an item is available for sales since the first week it had recorded sales until the last. This term accounts also for the regularity of the product sales. Full sales percentage on the last term compares the number of sales of the item made with regular price with its total sales, considering promotional and clearance sales. The three components were added as it would be more intuitive for the user to simply order the ratios to get the best items and it would consider three different characteristics on which the product had shown good performance.

This ratio is a hypothesis to measure product attractiveness. However several parameters could be considered and tested in order to make this sort of measurement. A suggestion for future work is given in order to experiment each of these components as an individual ratio and trying to create different ways to account for how good a product is for the customer so it can become the most complete measure. This sort of ratio can be different according to the data provided to analyze, being therefore adaptable for different companies.

4.1.8 Preprocessing set of tasks

The flowchart in Figure 15 shows the different steps taken for preprocessing data previously explained, describing how to go from raw data to the dataset presented in the beginning at Table 1. This process is a crucial phase on the overall methodology since it is where a deep understanding of data significance, context and relevance is made. Preprocessing knowledge extraction comes not only from the data but also from getting to know Company

X's business, how they take decisions regarding assortment, promotions and other product management matters. An alliance between data and the inputs from experienced colleagues on the business field allowed to reach a better understanding of some behavior reflected on the data. Besides that, some research was performed on the different brands' websites and social media pages in order to understand each brand concepts, price ranges and target customers. Events that happened in Mexico, holidays and festive epochs required investigation as well to understand some peaks of sales. The flowchart shows how all steps described are connected for the final purpose of data preparation.

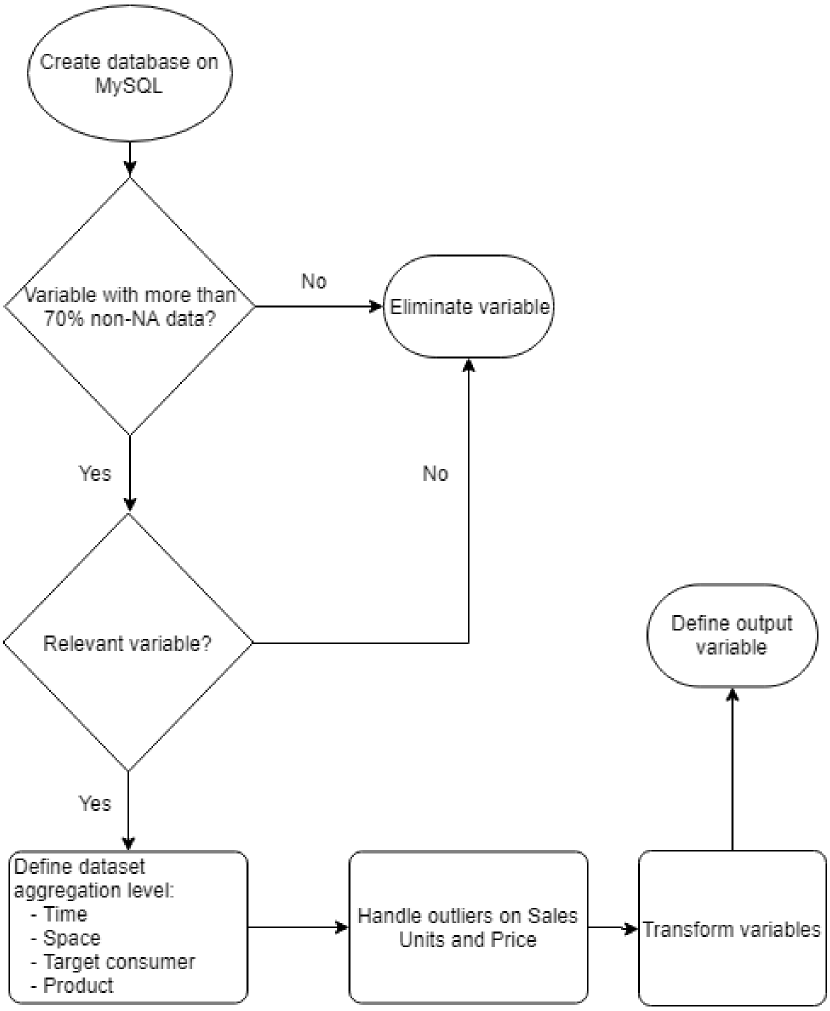


Figure 15 – Preprocessing flowchart

4.2 Statistical analysis over criteria selection – Tableau/SPSS

This section aims to give a more visual understanding of how product attributes and sales are related. For this purpose, Tableau software and Spyder were used to see how each variable would relate to sales and to each other.

4.2.1 Predictive variables and Sales Units

A first overview of sales distribution across each individual predictive variable was done through the construction of seven boxplots present in Appendix A.

Starting by sales type, as seen before in the outlier analysis and now in Appendix A Figure 1, it was possible to distinguish regular sales as the prominent type in both years, which provides a positive feedback to the Company X, as it means products recorded a higher number of sales units when they were being sold in full price more often. Despite that, looking at the price label boxplot in Figure 2 from Appendix A, it was possible to see most products have what are considered low and low medium prices. This makes sense since Company X usually applies low prices to their product by default.

Regarding the product type and family, these categories can be seen in Appendix A Figures 3 and 4, and pants and t-shirts can be highly distinguished as the key selling products. This was reinforced by product type sales records, evidencing sales from tops, bottoms and pants_2. Data showed to be coherent since product family records as they were included in the three types mentioned.

Looking at the gender in boxplot from Appendix A Figure 5 it was possible to notice a few products registered higher sales for women, however, the majority of items sales distribution seemed to be identical for both man and women.

Fashionability in Appendix A Figure 6 aims to measure the product's trendy characteristic and basics showed the highest recorded sales. Right after, the fashion basics were the second most sold.

Color was one of the attributes that had a bigger variety of possible values, seen in Figure 7 at Appendix A. It was possible to see a distinguishable pallet of basic colors under clients' most frequent purchase, namely black, white, grey and denim. Afterwards, contrasting with the first few, pattern/multicolor tones were the most sold.

Analyzing the brands in Appendix A Figure 8 it was possible to confirm that brand A was the strongest on sales, however, as it was classified in three components, the component A1 is the one the really takes over, followed by brands C and B.

In order to get a few more insights about how sales were characterized in the two years of study, an overview over how stores location could impact them was also performed. It was possible to confirm that from one year to the other little variation happened regarding regional comparison of annual sales, although the highest recorded value increased 2 million units in 2018. In terms of geography, sales along all North, West and South East maintained similar values with a total around 2 or 3 Million units. Center South is a remarkable region as it recorded the highest values of sales, contrasting with Southwest area that in both years registered the lower sales values of the country.

4.2.2 Crossed analysis on sales and predictive variables

After looking at variables individually, a deeper level of visual analysis was performed trying to breakdown predictors relationships that could be impacting for increasing sales. An iterative process was performed looking at combinations of variables to see which ones could bring out relevant and valuable information. The relevant sets of inputs were selected and explored in this section. Some marketing relationships were already known and confirmed through this analysis, for instance, certain products specified for one gender only.

Gender was firstly crossed with brand and product type and it allowed to visually confirm some products and brands that were only specified for either women or man. Figure 16 shows that sets are products only directed to women, whilst pants_2 includes only man's articles. Besides that, Figure 17 showed Brand A3 and C target consumer were women. Regarding gender, fashionability was analyzed, and it was observed a similar behavior across gender for all kinds. When it comes to color choice, women showed to have higher demand of denim comparing to man. The remaining colors had similar purchase registered.

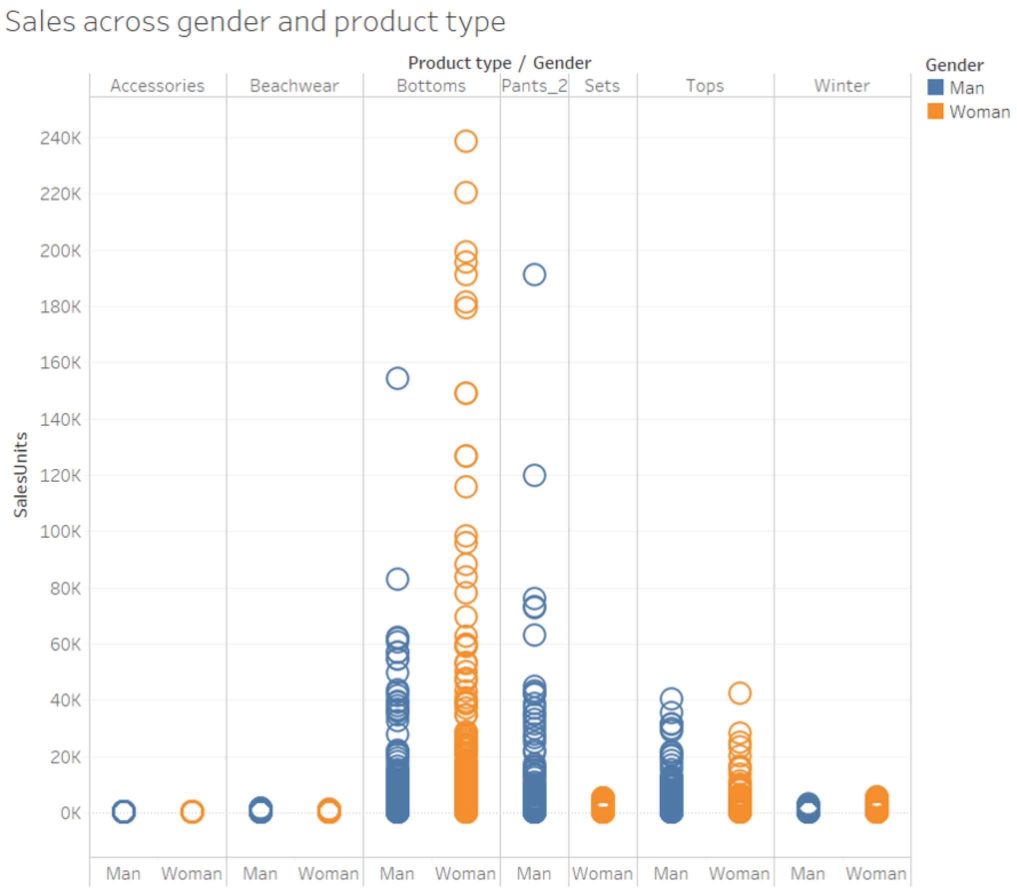


Figure 16 – Sales distribution across gender and product type

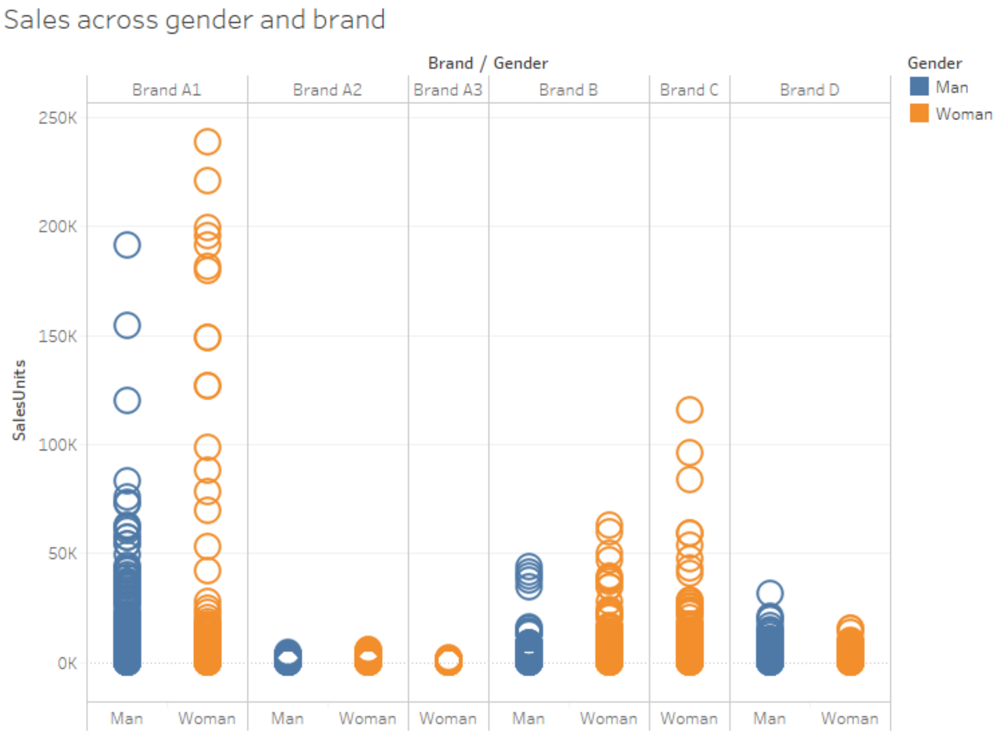


Figure 17 – Sales distribution across gender and brand

Price can be many times one decisive factor for a client to buy or not a given item. Therefore, it was decided to explore how it would vary along sale type, since, by definition, sale type varies due to a reduction in price or a promotional event which may be associated to it as well. Fashionability was included in this analysis as well, for matters of understanding if its value could be conditioned to different price ranges.

From Figure 18, it was possible to confirm that fashion items are the ones that suffer higher price fluctuations and yet, still catch client's interest, despite a way lower number of sales was registered for higher priced products.

One factor that may have influenced the observed values in Figure 18 is how price range variable was constructed. All products were considered when the different price ranges were defined. Since very few products had higher prices, this lead to building bigger intervals of price which were assigned a larger amount of products, while having very little number on the highest ones. Thus, the differences between prices don't seem to have much effect on sales as most products lie on the lower prices ranges. A suggestion for creating price labels by some product category is left in order to better understand price impact on sales.

Curiously, high priced items showed no register on promotional sales, however they were sold in clearance. These appeared to be very specific cases, given the little number of items within this price range. All types of sales showed to have a higher volume in medium low and low price products comparing to the remaining price labels, where a big difference can be seen regarding the order of magnitude of the respective scales. As the price increased, all types of sales decreased which shows the target consumer was probably a price-sensitive buyer. Basic products were clearly predominant in all sales types for low and medium low prices.

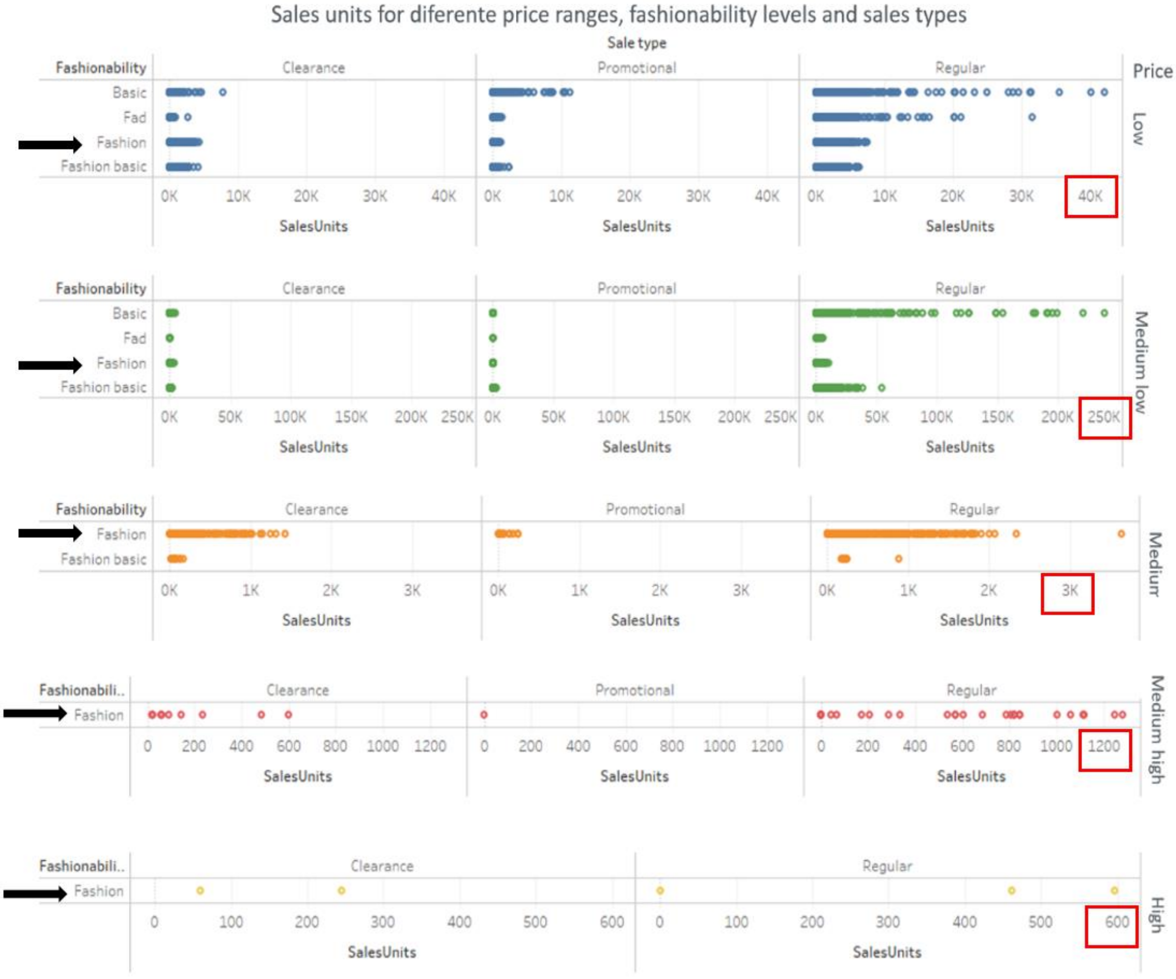


Figure 18 – Sales distribution across price range, fashionability and sale type

Another look was taken at brand and fashionability. It can be seen in Figure 19 that Fad products were a classification specified for Brand D. Besides that, all brands showed every fashionability value, except for Brand A3, which made sense since it is part of one out of three components of Brand A. Regarding these conclusions, for a specialist in Company X this would already be acquired knowledge, as the person would probably be familiar with these brand characteristics. However, it allows to see how the methodology put into practice easily allows for pattern recognition.

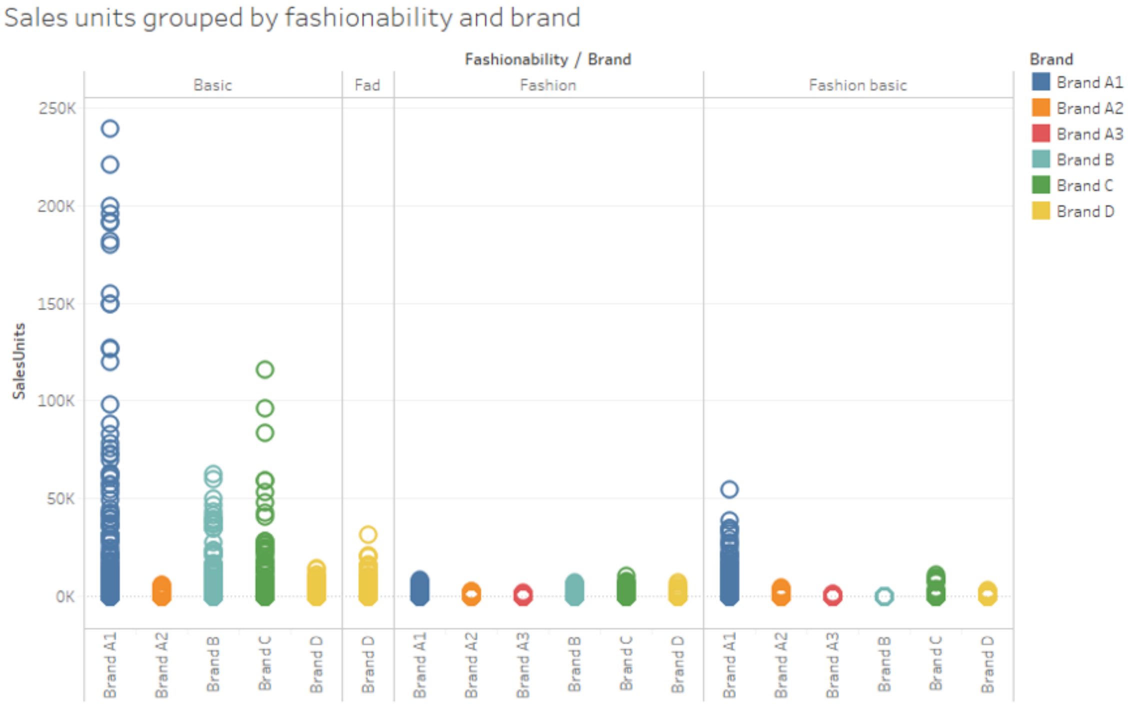


Figure 19 – Sales distribution across fashionability and brand

Product functionality can also lead to valorization of different attributes by the customer, so an analysis was performed on product type and family sales across fashionability, color and brand. Figures 20 and 21 showed that Brand D was the only selling accessories, swimsuits, beach complements and short sets. Looking at the graphs it was possible to confirm Company X's key selling product – pants. Brand A showed to have high sales as well on T-shirts. As for the remaining products, they all recorded relatively similar values, even though components 2 and 3 from Brand A have registered sales in a smaller variety of items, probably because they were a more specified range of product.

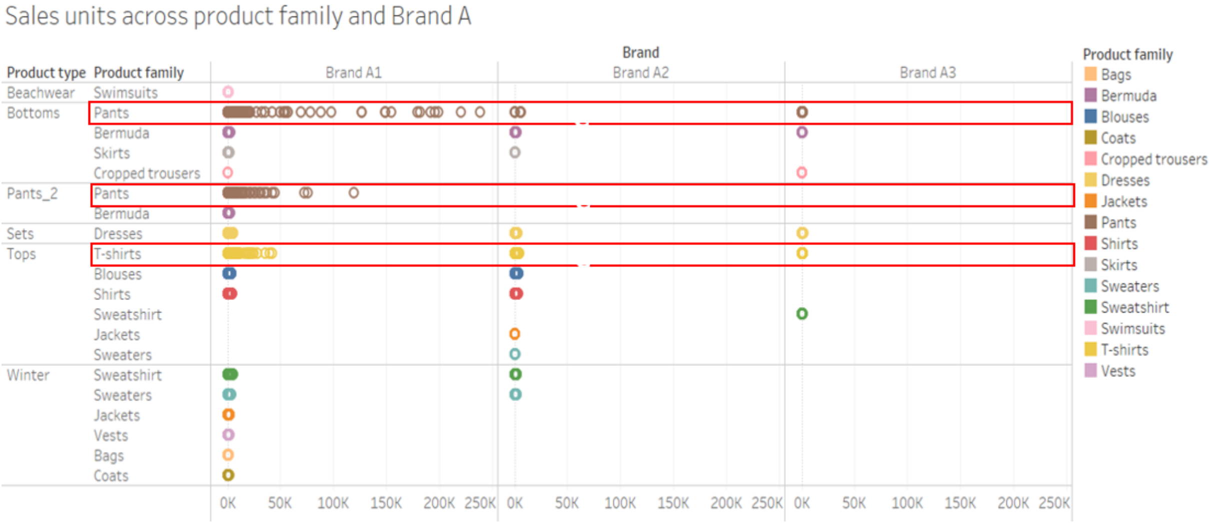


Figure 20 – Brand A sales distribution across product family

Regarding brands B and C, they showed very similar sales patterns across products having Bermuda with a few more sales for Brand C. Brand D appeared to be the widest brand either in terms of product family variety but also for having a specific kind of fashionability value. T-shirts were an item with higher volume of sales, followed by pants and skirts.

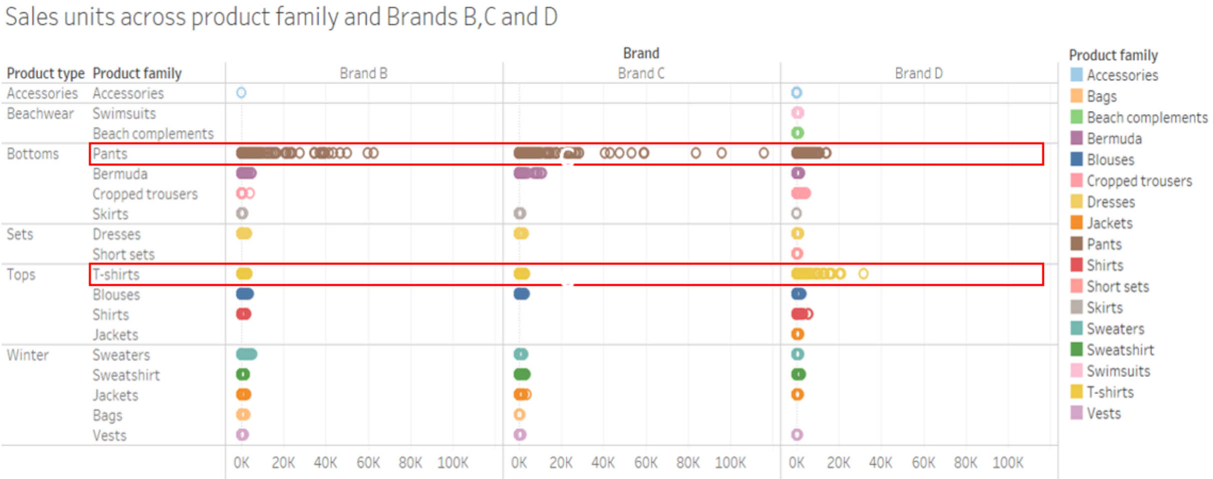
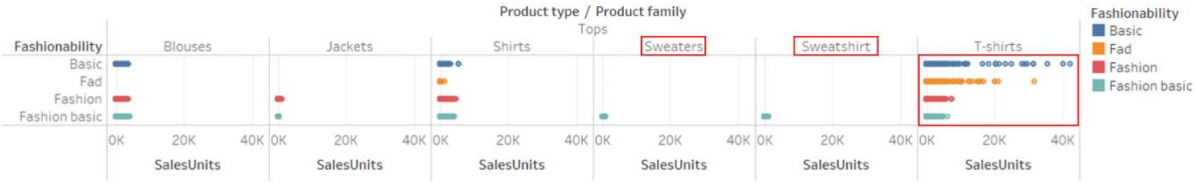


Figure 21 – Brands B, C and D’s sales units across product family

Observing product families and respective fashionability in Figures 22 to 25, some associations were suggested. For instance, in winter type of product sales, all product families registered to be fashion items, being sweaters and sweatshirts the only two product families with other fashionability levels. Fashion products were also the higher sales volumes in accessories, beachwear and sets. On the other hand, the types including the highest registered sales remained mostly in basic items, namely bottoms. Tops seem to have shown the widest sales with respect to their fashionability, since there were many basics, but also fashion and fashion basic items.



SalesUnits for each Fashionability broken down by Product type and Product family. Color shows details about Fashionability. The view is filtered on Product type, which keeps Tops.

Figure 22 – Tops sales units across fashionability and product family

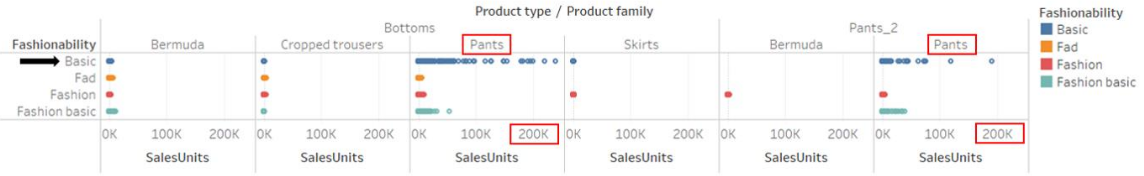


Figure 23 - Bottoms sales units across fashionability and product family

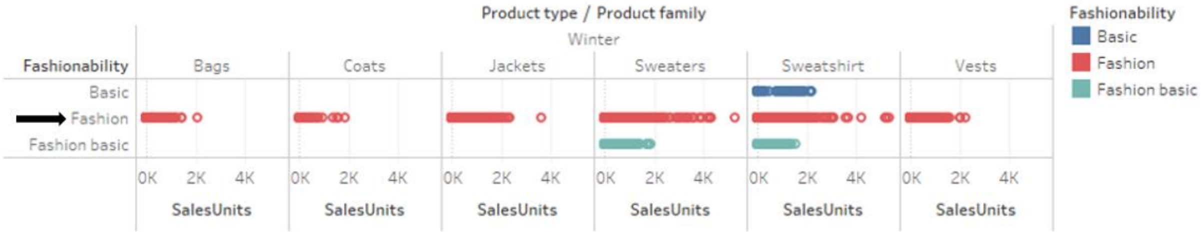


Figure 24 - Winter sales units across fashionability and product family

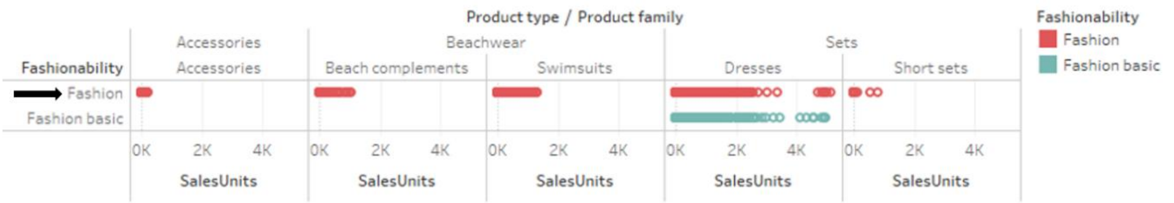


Figure 25 – Accessories, beachwear and sets' sales units across fashionability and product family

Color is an attribute that is strictly related to the physical appearance of garment, and according to Figure 26, neutral tones were the most demanded, namely white and black for blouses, shirts and T-shirts. However many other colors registered sales for this product type. Pants is a kind of item that is usually worn with many different sets of clothes and it was reflected on the colors sold, most of them being denim, black, blue and white as seen in Figure 27. Curiously, pattern/multicolor were also a pallet under client's eyes. The remaining product families showed sales in a wide variety of colors none of them standing out in a very contrasting way to the remaining, which can be seen in Figures 29 to 31. Despite that, black and multicolor items were a bit above the average.

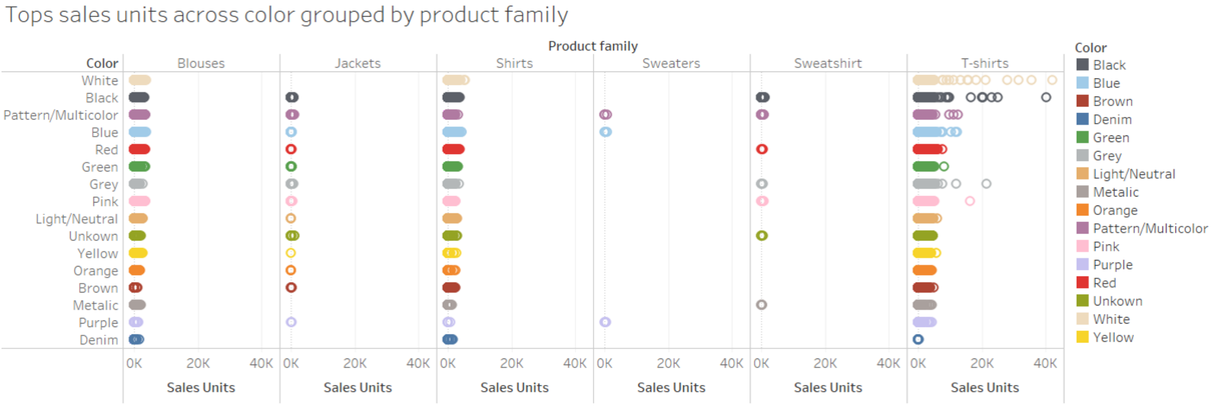


Figure 26 – Tops sales units across color and product family

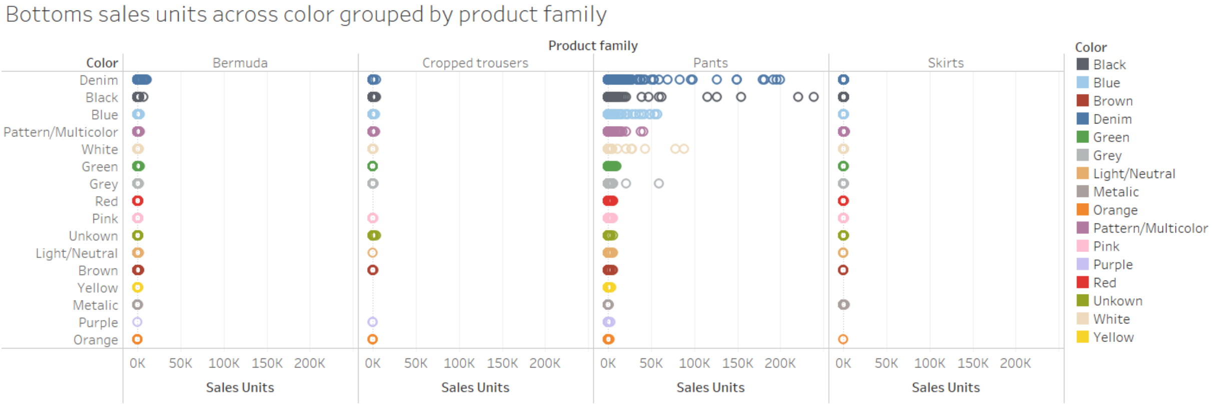


Figure 27 – Bottoms items sales units across color and product family

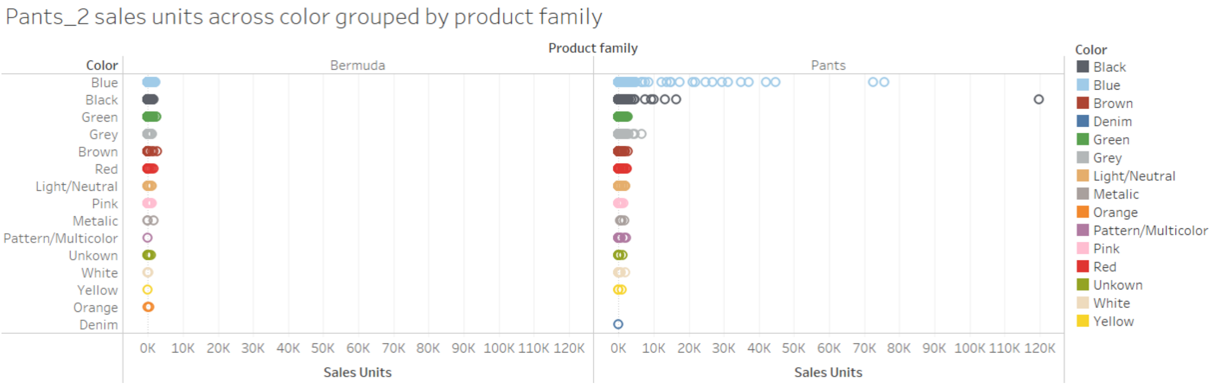


Figure 28 – Pants_2 sales units across color and product family

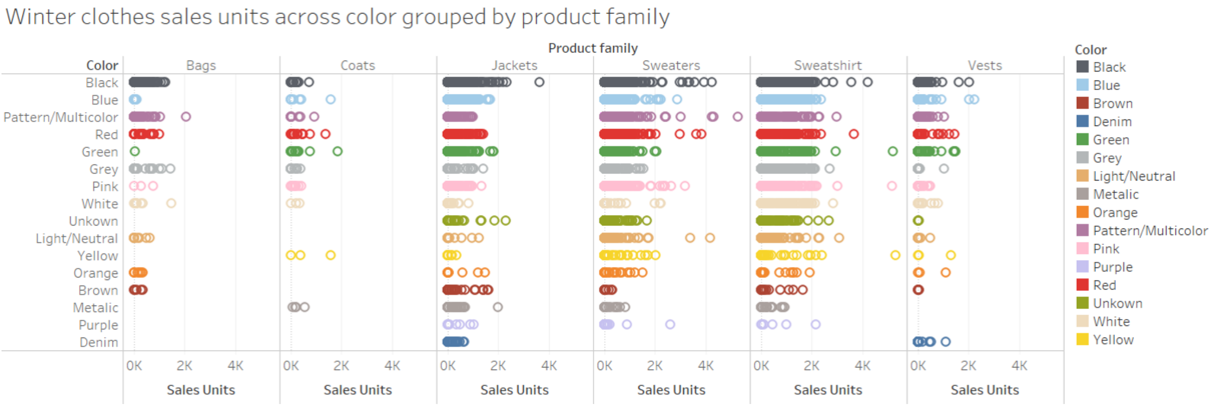


Figure 29 – Winter sales units across color and product family

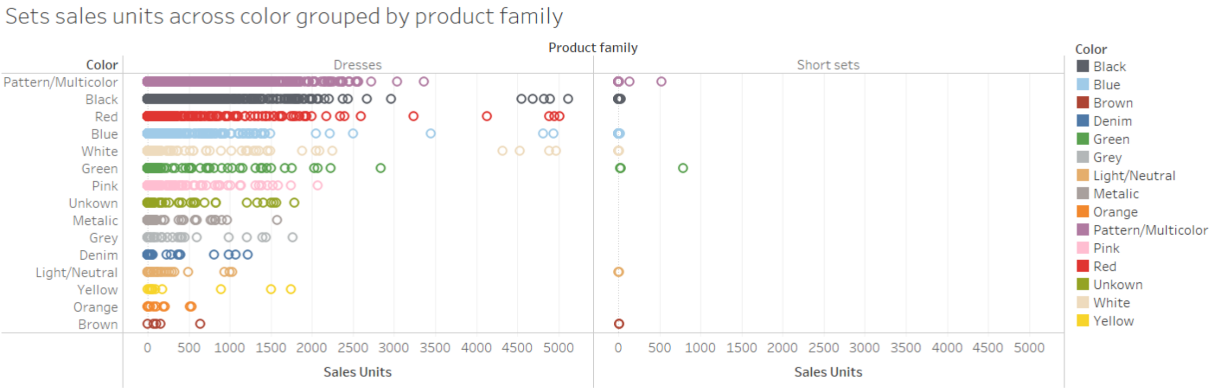


Figure 30 – Sets sales units across color and product family

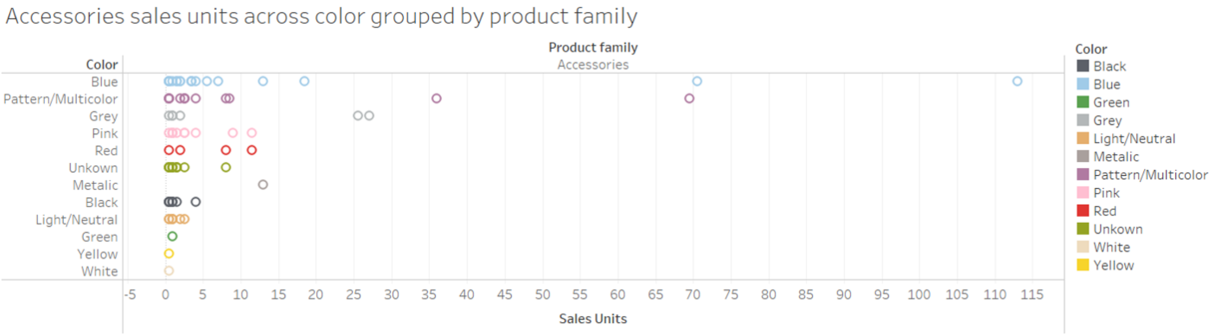


Figure 31 – Accessories sales units across color and product family

4.3 Predictive models – construction and results

4.3.1 Data splitting methods

For building the predictive models, Random Forest and Artificial Neural Networks were performed on RStudio, and Catboost was performed in Spyder environment. It will first be explained the logic behind the data splitting, training and testing, and afterwards the parameters chosen for each model.

This approach was common to all algorithms. Firstly, data was split into two subsets; 2017 observations were used for training and the following year for testing. Building the data mining models was a three steps process - parameter tuning, training the model and testing it.

A K-fold cross validation approach, with 10 folds, was taken for splitting data for parameter tuning. Using the training dataset, this sampling method separated data into ten equal folds and used nine of them for training and one for validating, switching the validation fold at each iteration. For every iteration, measures of performance mentioned in section 2.3.4 were computed comparing the predicted values of the validation set to the observed ones. Ten iterations were run for each parameters combination that was being tuned, saving a mean value of each of these error measures per parameter configuration. After running all possibilities, the set of parameters that registered the lowest error was chosen, in this case MAE. The choice MAE was based on the fact that comparing to RMSE and MSE this measure is less sensitive to outliers. Considering the provided dataset it was possible to see a few products with much higher sales than most of the remaining, so in order not to assign a higher weight to those observations RMSE and MSE were not the chosen ones. Between MAE and MAPE, choosing either one would lead to a similar choice, so MAE was chosen due to its easy interpretability. All measures were registered for model comparison, since each one had its own advantage advantage.

After the parameter tuning, the model was again trained using the optimized parameters and finally tested with the data from 2018. The predicted values were compared to the real ones and the models' performance was evaluated.

4.3.2 Parameter tuning configuration

For the Random Forest algorithm, 1000 trees were chosen for building the forest through an iterative process where values of 100, 500 and 1000 were experimented. Resampling was done for each bag and the number of variables available for split at each node was tuned, corresponding most of the times to the total number of predictive variables. This may be due to the low number of independent variables. A parameter available in the package, named importance, was also used. It measured the percentage of MSE increase if a predictive variable was left out of the model. This measure is very important since it defined priority on attributes analysis.

In Artificial Neural Networks, data had to be transformed because it did not accept non-numeric inputs. Thus, all categorical variables were transformed into dummy variables through one-hot encoding process. This process converts each value from a categorical variable into a single binary variable with values of 0 or 1. Besides that, a maxmin normalization was performed in the output variable, to speed up the convergence process. Regarding the network configuration, it was chosen for it to be a single-layered network, using backpropagation as it is the most commonly used method. Two optimization functions were tested, the Rectifier Linear Units and the hyperbolic tangent. Tuning of the learning rate and of the number of neurons was performed. For the output layer it is common practice to use a linear function, thus it was adopted. Regarding the number of epochs, 100 were performed with a batch size of 32 samples.

This configuration showed results using a package in R named Keras. However, this package wouldn't allow to access variable's importance. Thus, different packages which had that function were experimented and once again tuned. However, a problem came up since the algorithm never converged. This was possibly caused by the fact that all input variables were categorical having high cardinality, leading to a big number of binary variables as inputs. A decision was then taken, assuming this might not be the most adequate model to handle this kind of features, so Catboost algorithm replaced was tested.

Catboost, as a version of Gradient Boost was an algorithm particularly fit to handle categorical features. It was then adopted and three of its parameters were tuned, namely, the number of iterations, the learning rate and the trees depth. Regarding iterations the values of 50, 100, 500 and 1000 were experimented. The learning rate was tested for 0.01, 0.05, 0.2, 0.5, 0.8 and 1. The trees depth was assigned values from the lowest to the highest acceptable value 6 to 16. Regarding the loss function, RMSE was chosen, making use of a Bayesian bootstrap, a bagging temperature of 1 and l2 leaf regulation of 1.

4.4 Models predictive performance

When studying the results, two important analysis had to be taken regarding the machine learning models. The first one relates to the predictive performance of the models, comparing their performance in different scenarions and aggregation levels. The second one regards importance values extracted from the models and its alliance with the visual analysis insights.

The metrics adopted to evaluate the models performances are measured in the same units as the target variable for every scenario, except for MAPE, that is obtained as a percentage. Therefore, Table 3 shows the range each of the variables is spread along, and their average value. Regarding scenario 1, it is important to retain that, despite having a big spread of values, 99% of the observations lie within the interval of 0 to 10.000 sales units. Scenario 2 has 93,6% of the observations with values between 0 and 218 sales units per week. As for the ratio, in scenario 3, 93,4% of the stylecolors are in the interval between 0 and 7 of ratio. In all scenarios, the frequency distribution of the output variable has a high concentration of values in the small values of the variable, and as it increases, their frequency reduces drastically. This suggests sales could be approximate to a negative exponencial or a lognormal distribution.

Table 3 - Order of magnitude of output variables across scenarios

Product aggregation	**Scenario**	**Output variable**	**Observations range**		
			Minimum	Maximum	Average
All products	1	Sales units	1	500000	1012
	2	Sales units per week	0	15100	62
	3	Ratio	0,042	459,330	3,180
Tops	1	Sales units	1	84300	907
	2	Sales units per week	0	3762	58
	3	Ratio	0,044	73,711	3,054
Bottoms	1	Sales units	1	477583	1762
	2	Sales units per week	0	10640	89
	3	Ratio	0,042	459,328	4,205
T-shirts	1	Sales units	1	84297	1005
	2	Sales units per week	0	3762	63
	3	Ratio	0,045	73,711	3,215
Pants	1	Sales units	1	477583	2221
	2	Sales units per week	0	15083	114
	3	Ratio	0,042	459,328	4,879

Table 4 shows the results obtained for all product aggregation tries, different scenarions and using Random Forest (RF) and Catboost (CB). Comparing the two models, it is possible to see Catboost had a better performance for all scenarios and all product aggregations levels, which is comproved through all metrics. For this motive, further comparisons refer only to the values obtained with Catboost.

Table 4 – Models' performance results

Product aggregation	Scenario	Model	Evaluation metrics			
			MAE	MAPE (%)	RMSE	MSE
All products	1	RF	1282,98	15041%	5868,03	34433792,00
		CB	1090,67	7626%	4245,27	18022331,76
	2	RF	68,45	3151%	227,08	51565,02
		CB	58,73	1864%	201,79	40719,45
	3	RF	1,92	104%	6,84	46,73
		CB	1,73	54%	6,78	45,93
Tops	1	RF	1002,10	11914%	1767,57	3124295,00
		CB	946,87	6591%	1705,55	2908909,53
	2	RF	62,87	3118%	106,71	11386,30
		CB	53,21	1677%	100,97	10194,25
	3	RF	1,59	87%	2,20	4,86
		CB	1,50	53%	2,13	4,53
Bottoms	1	RF	2072,23	17839%	10615,44	112687533,00
		CB	1890,63	9884%	10079,80	101602405,13
	2	RF	101,82	3924%	469,08	220035,90
		CB	81,69	1808%	386,18	149133,31
	3	RF	3,84	183%	16,43	269,82
		CB	3,52	109%	16,07	258,34
T-shirts	1	RF	1135,84	13377%	1982,21	3929171,00
		CB	1050,08	6901%	1922,57	3696273,06
	2	RF	68,44	3510%	115,57	13355,38
		CB	58,32	1765%	111,00	12321,99
	3	RF	1,76	96%	2,43	5,93
		CB	1,65	54%	2,37	5,63
Pants	1	RF	3211,14	33557%	15867,67	251782967,00
		CB	2519,84	16292%	10369,61	107528892,64
	2	RF	134,88	4840%	580,86	337399,40
		CB	111,36	2291%	505,46	255490,02
	3	RF	4,86	230%	18,26	333,48
		CB	4,30	123%	17,70	313,24

From an overall perspective, looking at all models across different scenarios, Ratio prediction scenarios show better results comparing to the sales units ones, considering that scenario 1 is the one presenting worse metrics.

When comparing MAE and RMSE with the mean values registered for output variables it is possible to see that the aggregation level revealing highest error predictions is bottoms, having for scenario 1 an RMSE which represents almost six times the average value registered for sales units in the observations. The worst values were all found for bottoms either, aggregating by product type, or by a more detailed level, in this case, pants. A possible cause for these results relates to the fact that pants are the item with the highest resgistered sales, having units of stylecolors that go from a range of 1 to almost 500 million units sold.

A similar behavior regarding the proportion of MAE and RMSE over the average output variable registries was found between tops and T-shirts. This may be due to the fact, that this is the product reccording the highest sales volume in the respective type of product.

When analysing predictions from product type aggregation level to product family, a decrease of error is found in both cases. As the level of detail within that group of products increases, it makes sence that the error reduces, as it probably has patterns more similar within the group making it easier for the algorithm to recognize them. When comparing lower aggregation levels with all products together, there are both increases and dicreases of error, so it is hard to take a conclusion out of it. In general, all measures of error appear to be very high considering the registered values from historical data.

4.5 Importance of predictive variables

This subchapter explores the importance assigned to variables extracted from each model, providing an example on how to apply them. As the models use different algorithms and were used from different packages in distinct working environments, their importance is computed in dissimilar ways. Random forest importance represents the decrease in MSE if the model is predicted without the variable. It was transformed into a relative percentage in order to make it easier to understand. When it comes to Catboost, features importance is calculated in a different way, importance represents "how much on average the prediction changes if the feature value changes" (Yandex 2019). This value is already provided by the package in percentage. Considering the models analyzed performances, Catboost importance was chosen to be analyzed. However, Random Forest results for importance can be accessed through Appendix B.

Several aggregation levels were experimented regarding the product to achieve these importance values. The choice of aggregation level is up to the analyst to choose and it can be made according to the kind of application he is using, the model best performance or the criteria that has a higher relevance according to the context.

An example of analysis will be provided on tops, assuming an aggregation level by product type. Figure 32 shows the importance values achieved for the chosen product type. The other importance results are provided in Appendix C.

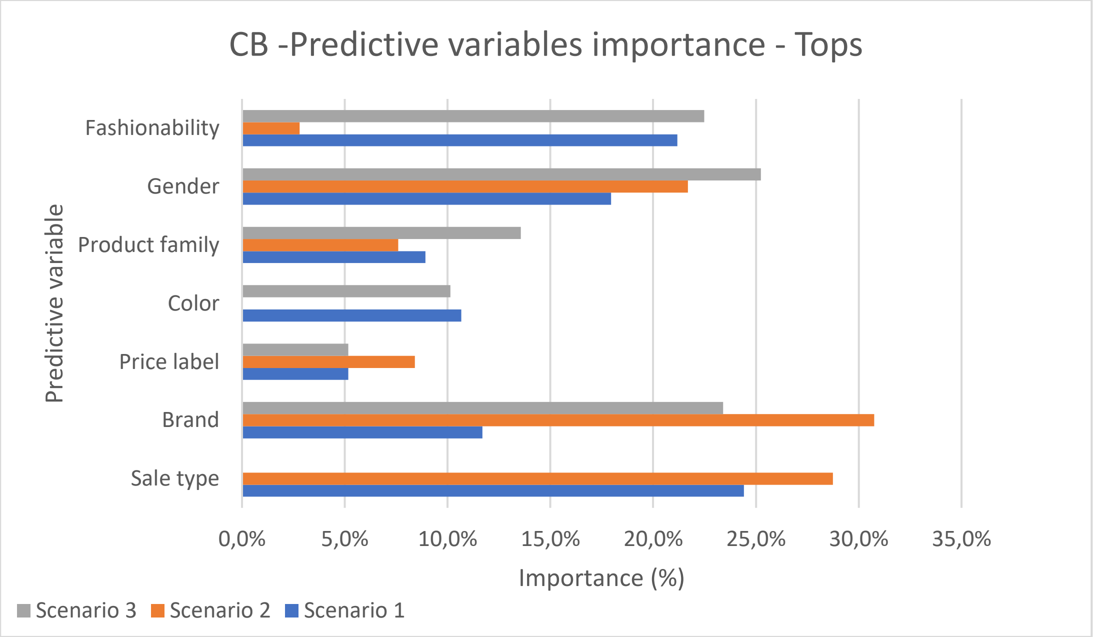


Figure 32 – Features importance results for tops

According to the results of the models explored before, an option was taken to go for scenario 3 importance results, since it had the lowest error indicators. Thus, the priority order for assortment planning of the following year for tops should become, according to the graph, gender, brand, fashionability, product family, color and finally price label. Considering this order, and allying the visual analysis previously made in section 4.2 and showed again in Figures 33 to 39, a suggestion would be given to choose for women most products of brand A1, B and C while for man, from brands A1, B and D. Figure 33 shows sales distribution across gender was not very different. Figure 34 shows inside the red square the brands that achieved sales higher than 30K units for some products, which led to the advice for choosing the mentioned brands for each gender.

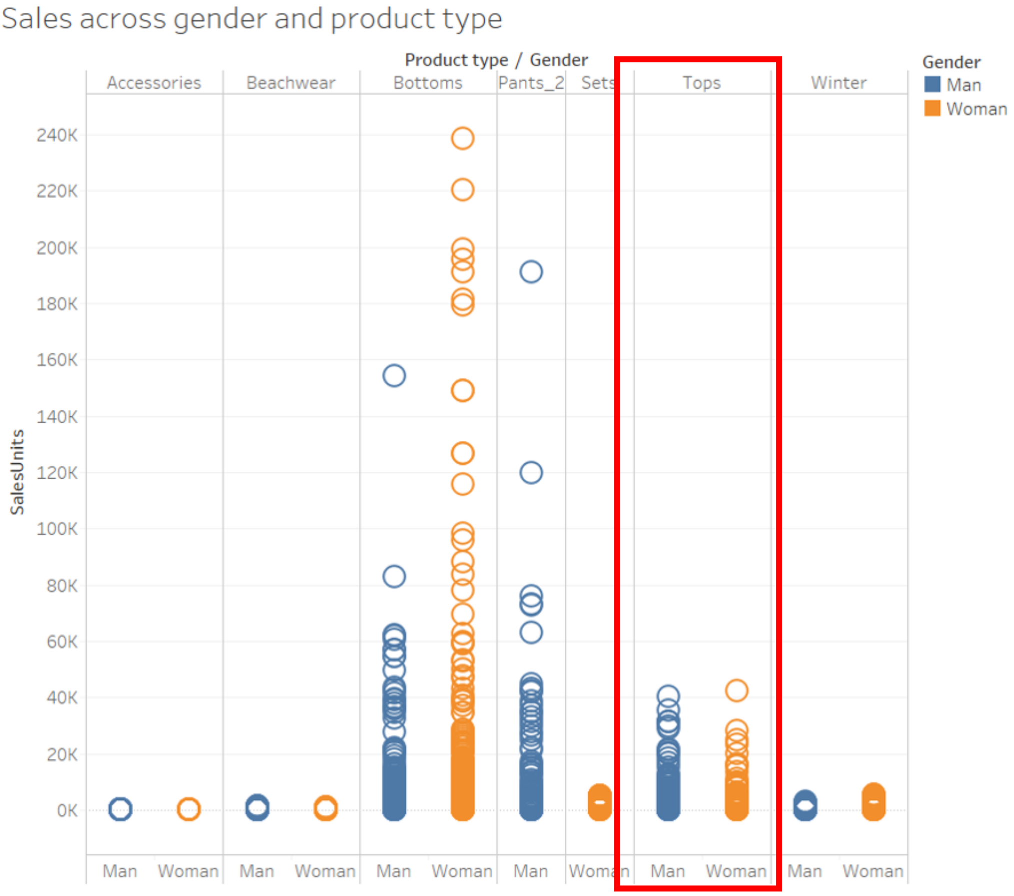


Figure 33 – Sales across gender and product type highlighting tops

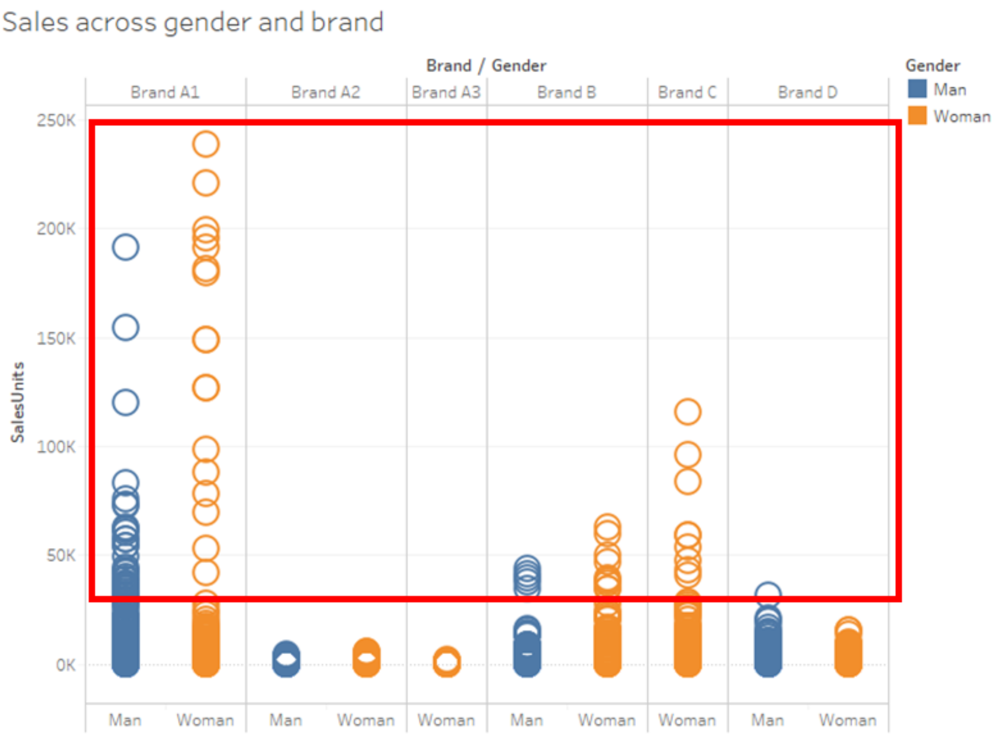


Figure 34 – Sales across gender and brand

Regarding fashionability, an advice to bet on a combination with a strong component of basic items and the remaining equally distributed would be given, by the analysis of Figures 35 and 36. In Figure 35 basic items showed to be the most purchased ones in regular sales, which are the kind of sale Company X wishes to maximize. The arrows in the figure point for sales scale when describing higher priced items that include no basic items. Given the order of magnitude of their scale, the key bet fashionability level remains on basic items. Figure 36 restricts fashionability to tops, confirming higher sales for basics. However sales were registered in the remaining levels of fashionability which led to an equal distribution choice on that sense.

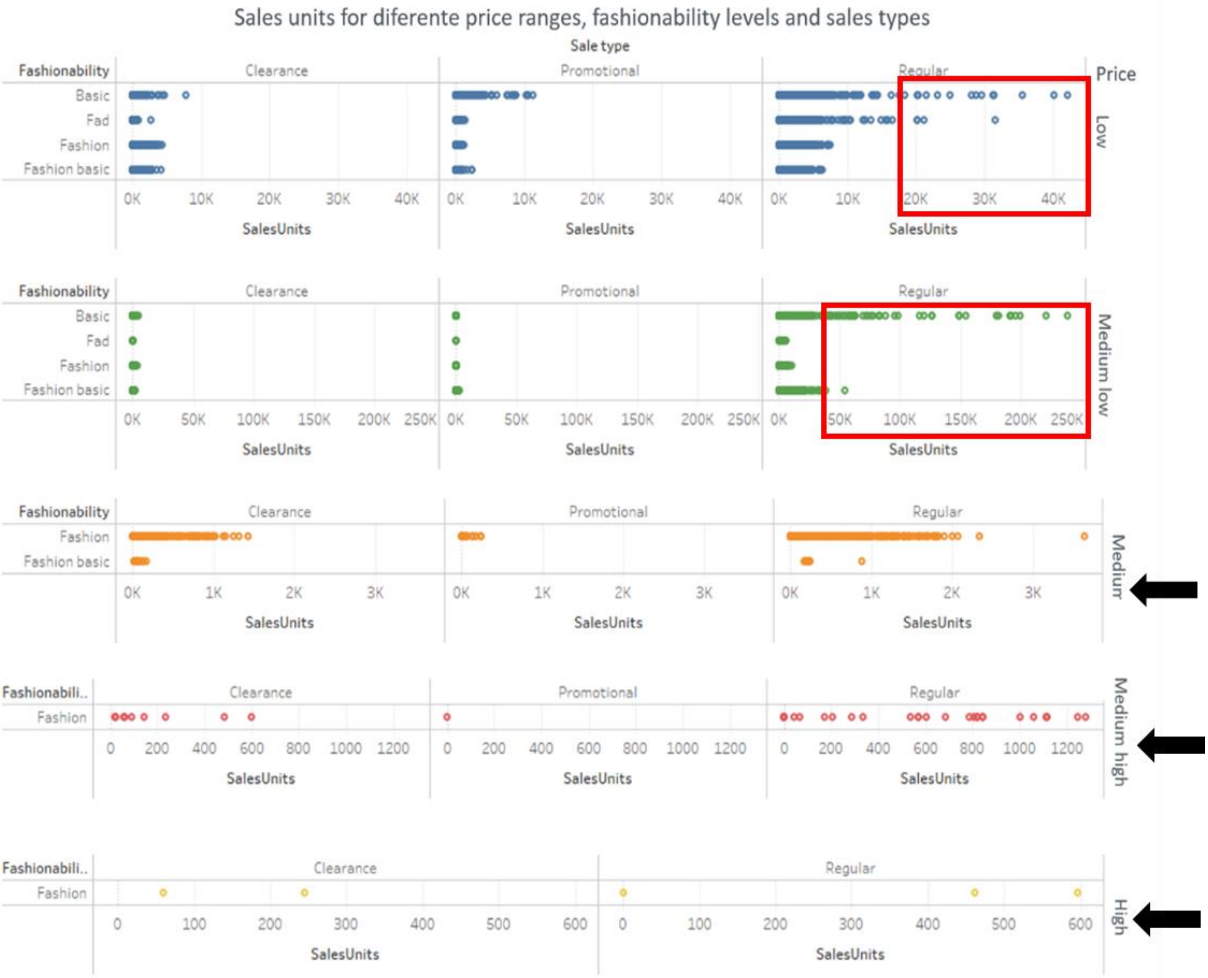


Figure 35 – Sales across price range, fashionability an sales type

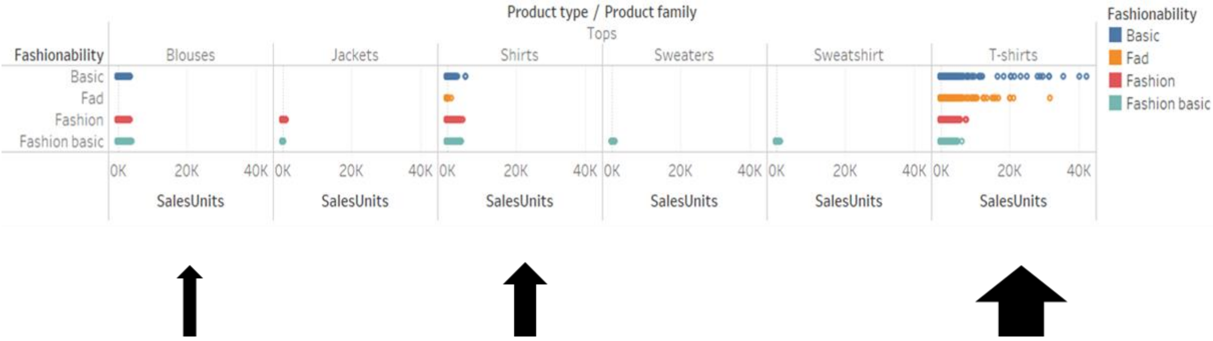


Figure 36 – Tops sales across fashionability levels

As for product family, a key bet are T-shirts. Shirts and blouses should be the second highest chosen products which is remarked by the proportion of arrows in Figure 36 and the red square on Figure 37. As for Figure 37 and 38, it shows that no matter the brand T-shirts, shirts and blouses are important product selections.

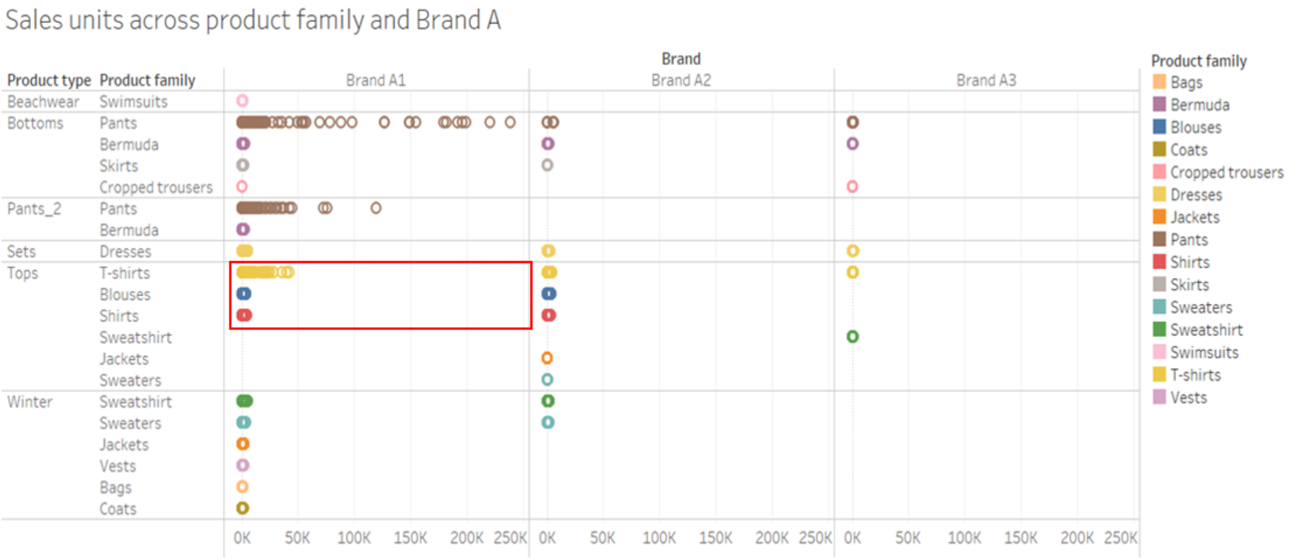


Figure 37 – Sales across brand A and product

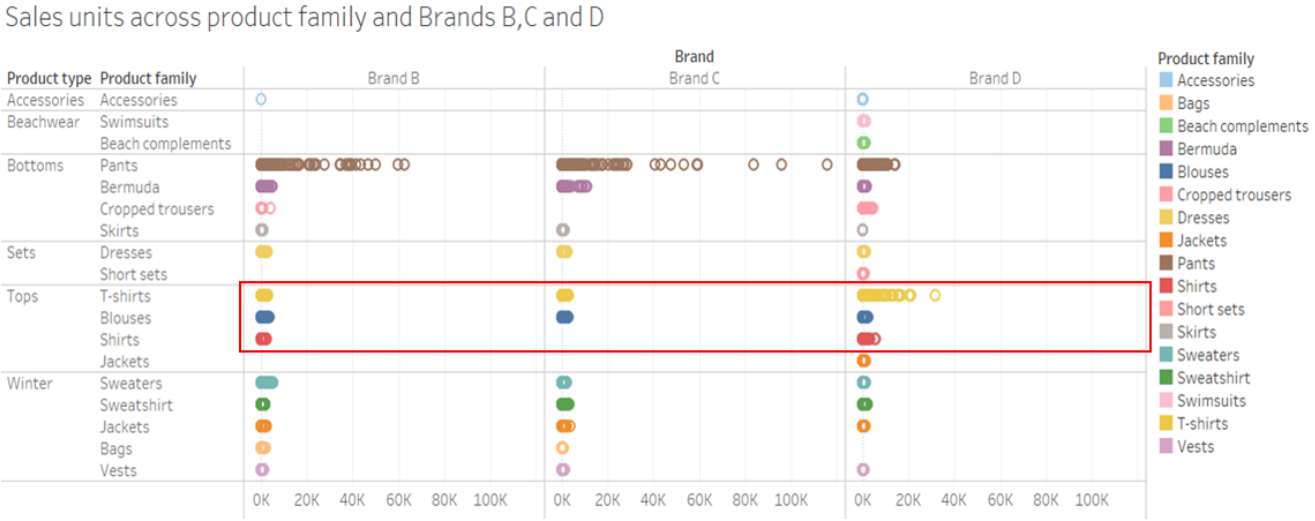


Figure 38 – Sales across brand and product

When it comes to tonalities, Figure 39 shows a big portion should go for black, white and blue items, followed by patterns and multicolor ones. The remaining should be equally distributed by the other colors in such a way that an equal amount of each color could be found on tops.

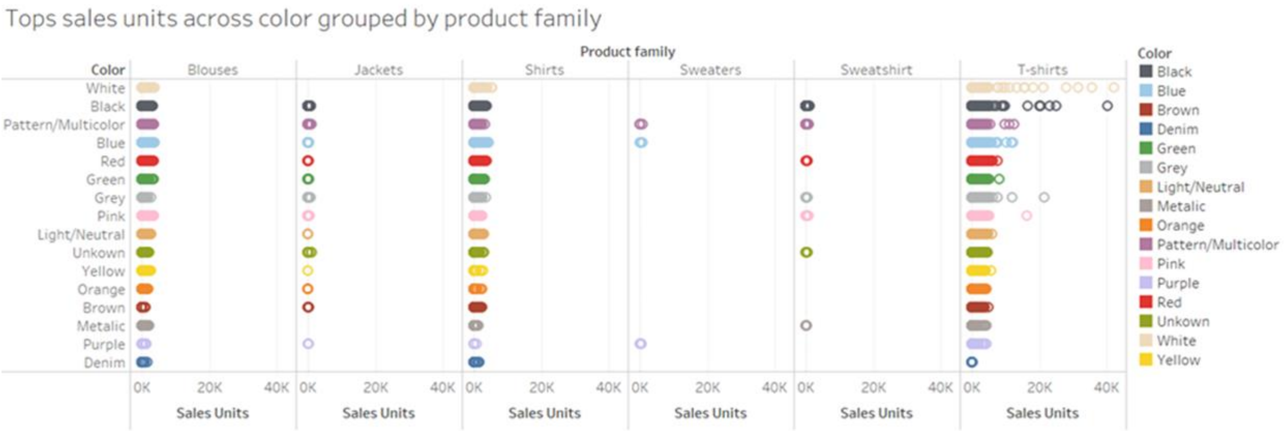


Figure 39 – Tops sales across color and product family

Supposing data about the new possible stylecolors to be selected for assortment would already be available, a third source of information for these criteria decision could come from introducing the data into the predictive model and predict their ratios. From those results, priority should be given to the items attributes that predicted to have higher ratios.

5 Conclusions and future work

This analysis is a suggestion that includes information mainly from historical data and the attributes importance assignment. It can be interpreted from several perspectives and it should be complemented with knowledge from other sources of the brand activity. An expert who is familiar with the market and new trends would be the ideal user to make this kind of decisions, as she or he would represent a more informed analyst.

Fashion is a very unpredictable market and trends are changing every year. Thus, the alliance of business, market and fashion knowledge with the methodology proposed in this dissertation would be an interesting way to make a more informed decision.

For future work, it would be useful to collect information from fashion experts who work on the brand and are familiar with the usual customer behavior as it can add valuable information to the methodology that is not always so easy to extract directly from the data. Other levels of aggregation and other algorithms for variables importance assignation should be tested.

Despite the unsuccessful attempt to run artificial neural networks, a binary input variable network was found on literature in (Dürichen et al. 2018) that could be explored, as it would probably be more adapted to the categorical data used for these predictions.

In general, the methodology presented can be very useful in the assortment planning process in the widest applications. Despite having results with relatively high error, this approach should be tried on other datasets to test the method robustness.

The proposed method could be relevant to facilitate visibility on finding certain correlations between attributes. It is important to remark, however, that when analyzing them, the causes for what is graphically demonstrated can be not only related to sales or the product itself, but also, by the way variables are built such as what happens in the case of price range. Thus, an extreme importance is assigned to the preprocessing phase, as the way variables are built can have a big impact on the analysis.

Quantitative measures were not obtained to weight attributes, however their relative importance and the best options to choose them can be taken from the method presented. The analysis provided has some subjectivity since no quantitative weights have been achieved for product characteristics. However, transforming the results of this analysis into percentages could be limiting the process to adding the additional knowledge of the analyst to the results. By giving a priority suggestion, the possibility of introducing new trends information for instance is allowed for the user.

References

Baecke, Philippe, Shari De Baets, and Karlien Vanderheyden. 2017. “Investigating the Added Value of Integrating Human Judgement into Statistical Demand Forecasting Systems.” *International Journal of Production Economics* 191 (June): 85–96. https://doi.org/10.1016/j.ijpe.2017.05.016.

Barbezat, Suzanne. 2018. “Mexico’s Weather: What to Expect.” 2018. https://www.tripsavvy.com/mexicos-weather-what-to-expect-1589004.

Choi, Tsan Ming, Chi Leung Hui, Sau Fun Ng, and Yong Yu. 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): 1003–10. https://doi.org/10.1109/TSMCC.2011.2176725.

Consult, Retail. 2015a. “Retail Consult|About Us - Our Team.” 2015. http://www.retail-consult.com/about-us.

———. 2015b. “Retail Consult|Our Partners.” 2015. http://www.retail-consult.com/clients-partners-2.

———. 2019. “Corporate Onboarding.”

Dürichen, Robert, Thomas Rocznik, Oliver Renz, and Christian Peters. 2018. “Binary Input Layer: Training of CNN Models with Binary Input Data,” no. Nips: 2–5. http://arxiv.org/abs/1812.03410.

Gong, Hongren, Yiren Sun, Xiang Shu, and Baoshan Huang. 2018. “Use of Random Forests Regression for Predicting IRI of Asphalt Pavements.” *Construction and Building Materials* 189: 890–97. https://doi.org/10.1016/j.conbuildmat.2018.09.017.

Goodman, Bernadette. 2016. *Oracle® Retail Assortment Planning User Guide*. *Oracle AP User Guide*.

Han, Jiawei, and Micheline Kamber. 2001. *Data Mining: Concepts and Techniques*. Edited by Academic Press. Morgan Kaufman publishers.

Ibrahim, Dogan. 2016. “An Overview of Soft Computing.” *Procedia Computer Science* 102 (August): 34–38. https://doi.org/10.1016/j.procs.2016.09.366.

Jegethesan, K.; Sneddon, J. N.; Soutar, G.N. 2007. “Young Australian Consumers’ Preferences for Fashion Apparel Attributes.” *Journal of Fashion Marketing and Management* 11 (4): 275–90. https://doi.org/10.1108/03090560410539302.

KDnuggets. 2019. “Understanding Machine Learning Algorithms.” 2019. https://www.kdnuggets.com/2017/10/understanding-machine-learning-algorithms.html.

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

Natekin, Alexey, and Alois Knoll. 2013. “Gradient Boosting Machines, a Tutorial.” *Frontiers in Neurorobotics* 7 (DEC). https://doi.org/10.3389/fnbot.2013.00021.

Ni, Yanrong, and Feiya Fan. 2011. “A Two-Stage Dynamic Sales Forecasting Model for the Fashion Retail.” *Expert Systems with Applications* 38 (3): 1529–36. https://doi.org/10.1016/j.eswa.2010.07.065.

Shoham, Aviv. 2008. “Determinants of Fashion Attributes ’ Importance.” *Journal of International Consumer Marketing* 1530 (January 2014): 37–41. https://doi.org/10.1300/J046v15n02.

Swalin, Alvira. n.d. "NoChoosing the Right Metric for Evaluating Machine Learning Models — Part 1." Accessed June 21, 2019. https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4.

Tech, Georgia. 2016. "Boosting - Udacity." 2016. https://eu.udacity.com/course/machine-learning-for-trading--ud501.

Thomassey, Sbastien. 2010. "Sales Forecasts in Clothing Industry: The Key Success Factor of the Supply Chain Management." *International Journal of Production Economics* 128 (2): 470–83. https://doi.org/10.1016/j.ijpe.2010.07.018.

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

Thomassey, Sébastien, and Michel Happiette. 2007. "A Neural Clustering and Classification System for Sales Forecasting of New Apparel Items." *Applied Soft Computing Journal* 7 (4): 1177–87. https://doi.org/10.1016/j.asoc.2006.01.005.

Wong, W. K., and Z. X. Guo. 2010. "A Hybrid Intelligent Model for Medium-Term Sales Forecasting in Fashion Retail Supply Chains Using Extreme Learning Machine and Harmony Search Algorithm." *International Journal of Production Economics* 128 (2): 614–24. https://doi.org/10.1016/j.ijpe.2010.07.008.

Xia, Min, Yingchao Zhang, Liguo Weng, and Xiaoling Ye. 2012. "Fashion Retailing Forecasting Based on Extreme Learning Machine with Adaptive Metrics of Inputs." *Knowledge-Based Systems* 36: 253–59. https://doi.org/10.1016/j.knosys.2012.07.002.

Yandex. 2019. "Feature Importance." 2019. https://catboost.ai/docs/concepts/fstr.html#fstr__regular-feature-importance.

APPENDIX A: Attributes vs Sales units boxplots

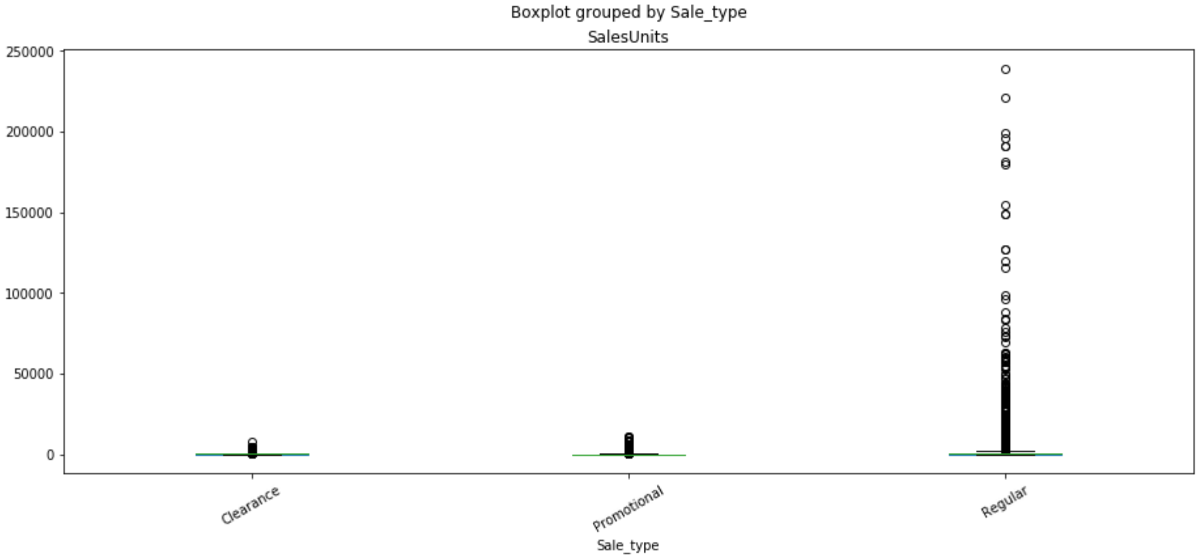


Figure 1 – Boxplot of sales units grouped by sale type

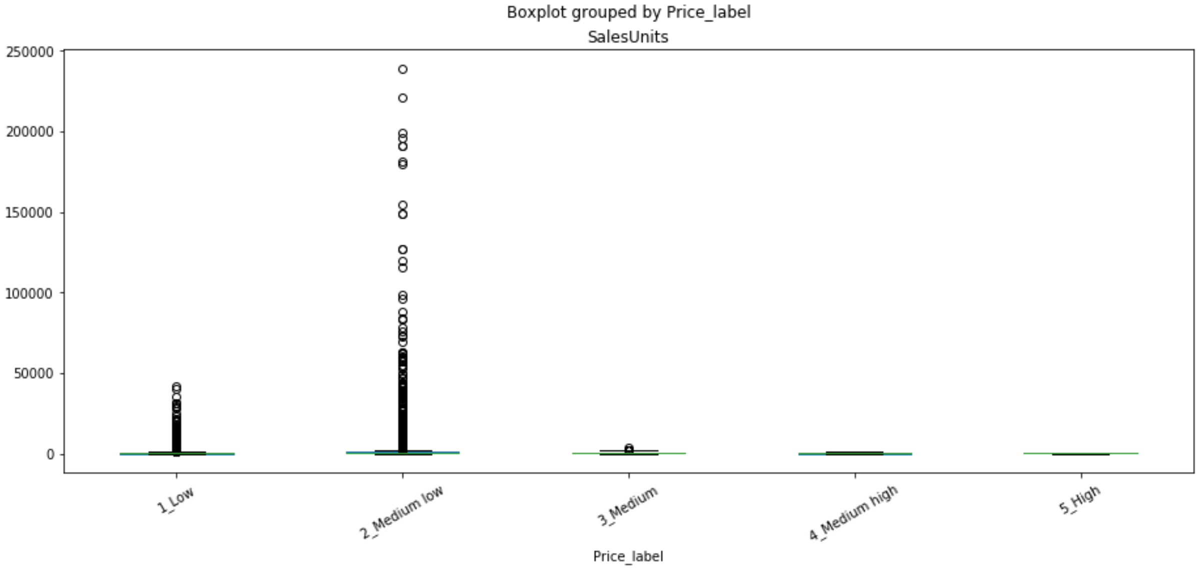


Figure 2 - Boxplot of sales units grouped by price label

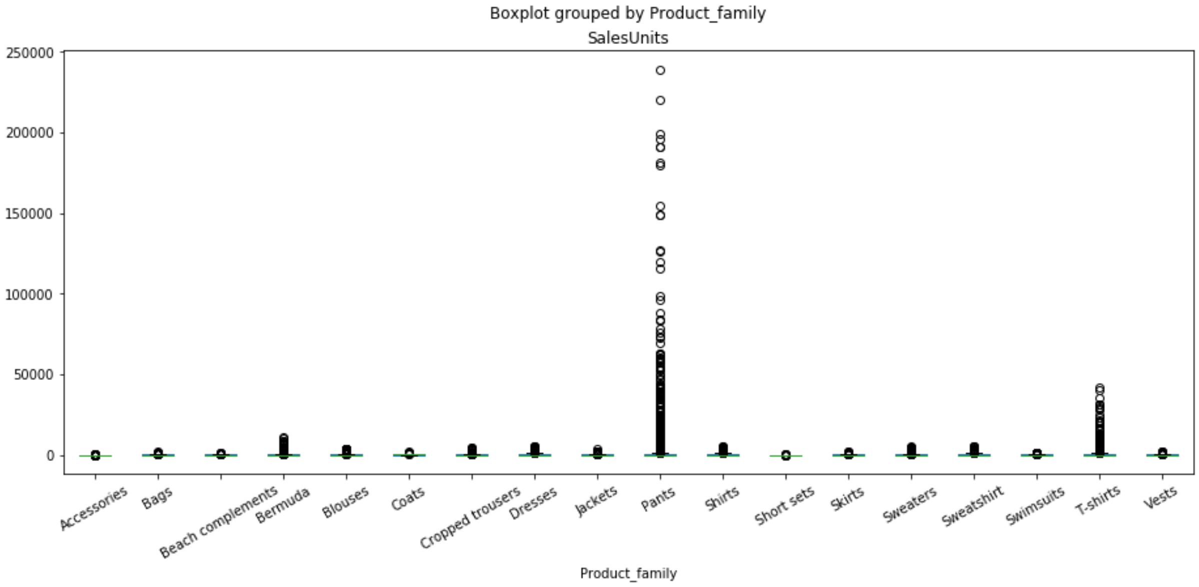


Figure 3 - Boxplot of sales units grouped by product family

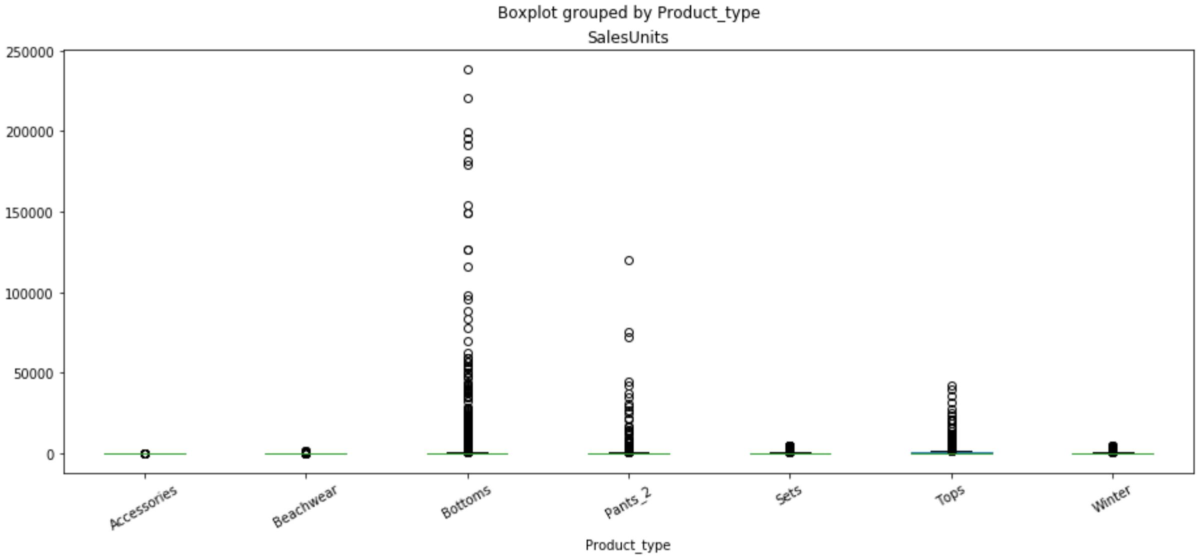


Figure 4 - Boxplot of sales units grouped by product type

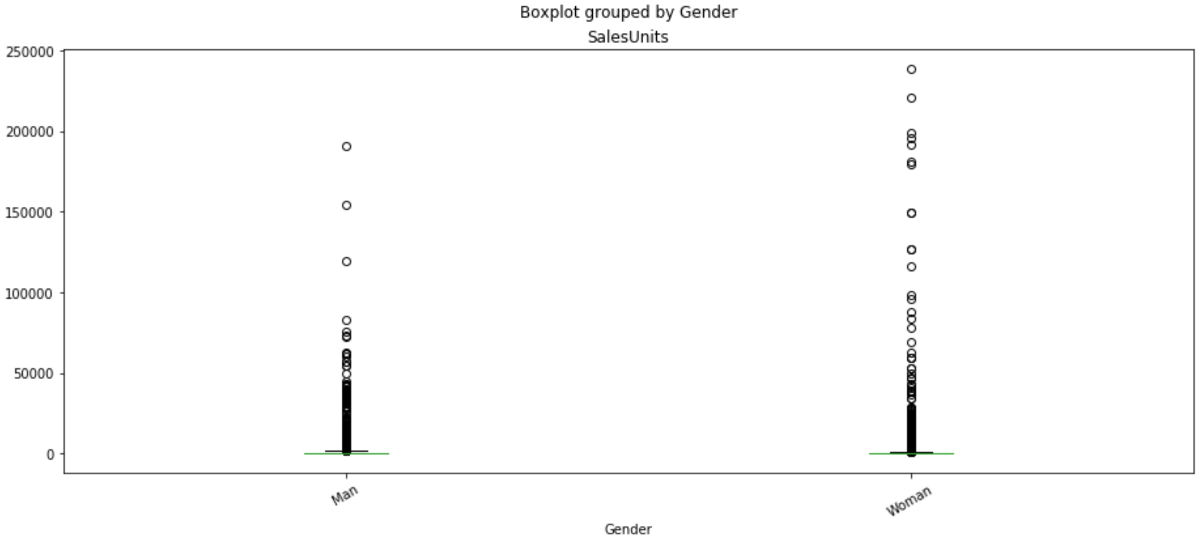


Figure 5 - Boxplot of sales units grouped by gender

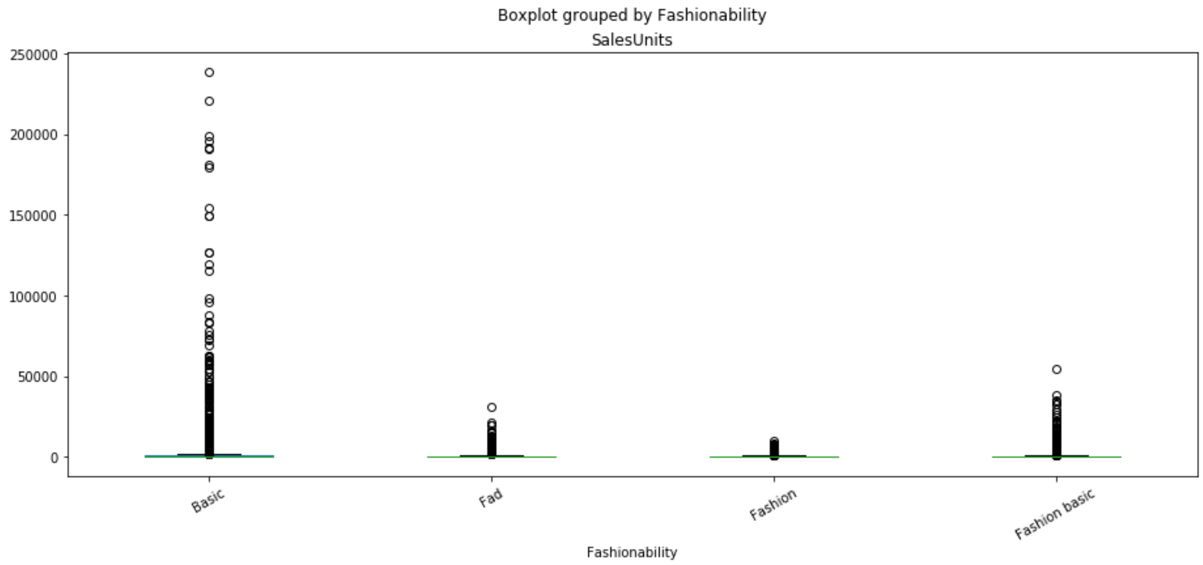


Figure 6 - Boxplot of sales units grouped by fashionability

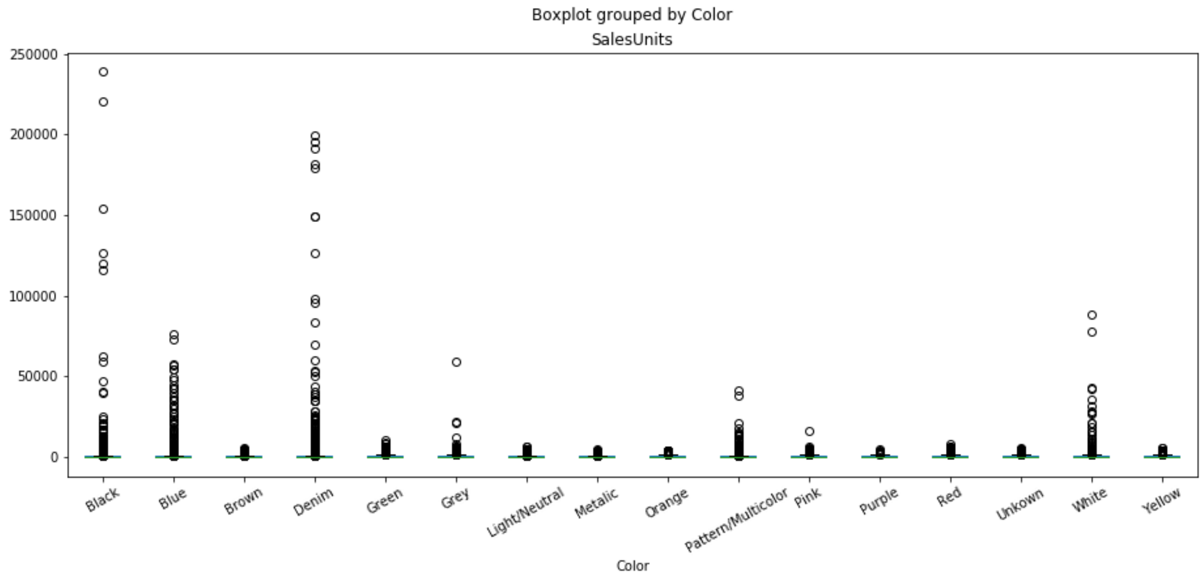


Figure 7 - Boxplot of sales units grouped by color

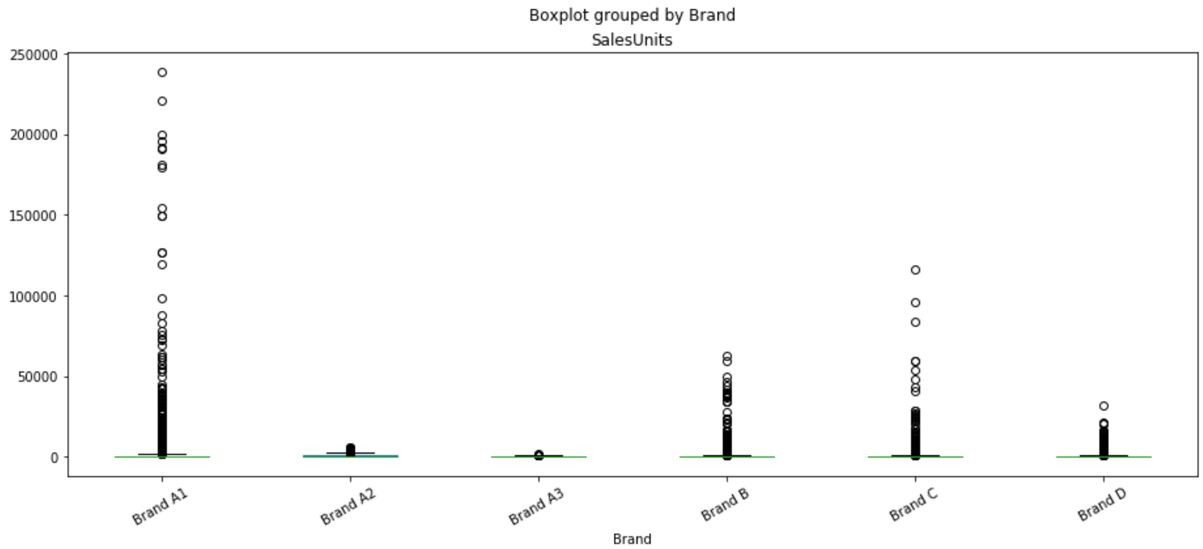


Figure 8 - Boxplot of sales units grouped by brand

APPENDIX B: Random Forest features importance results

Considering:

- Scenario 1 – Units sold
- Scenario 2 – Units sold per week
- Scenario 3 – Ratio

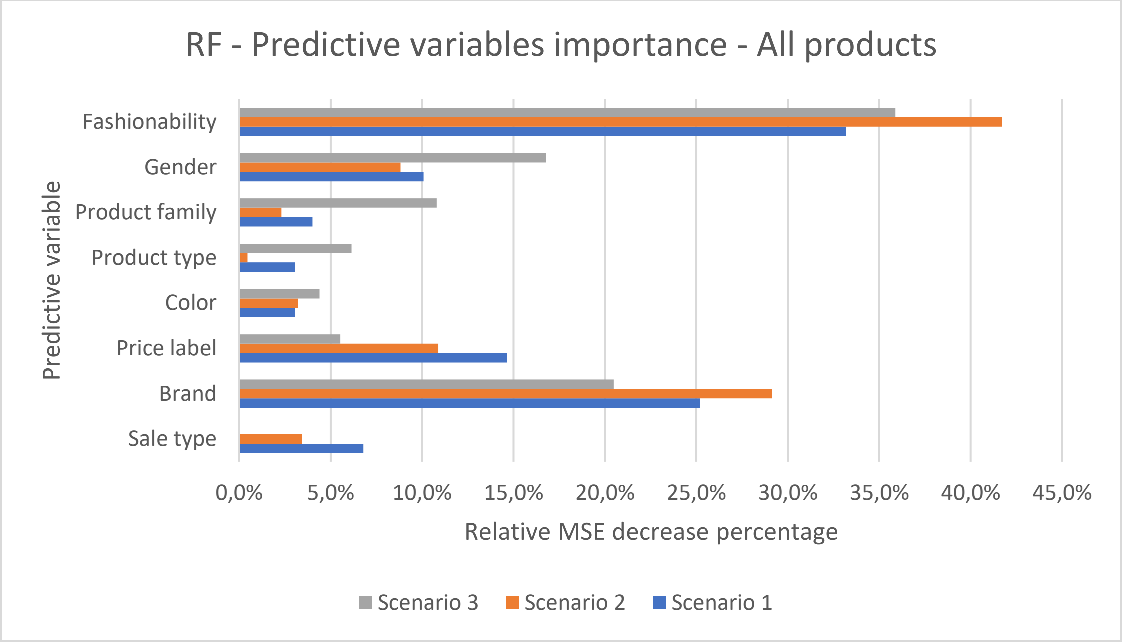


Figure 1 – Features importance results for all products

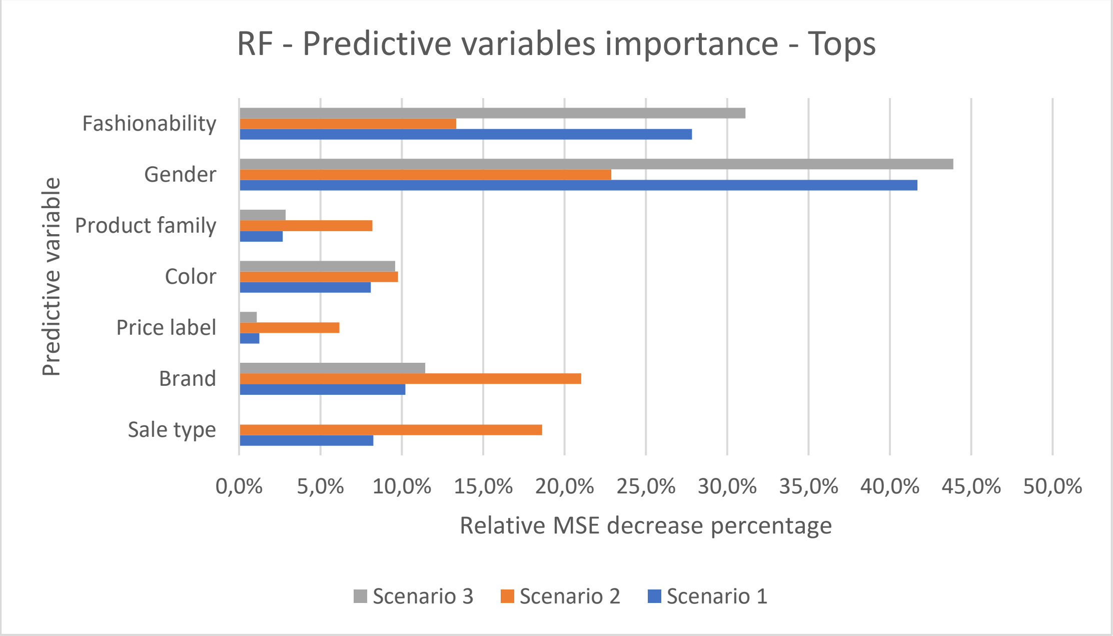


Figure 2 – Features importance results for tops

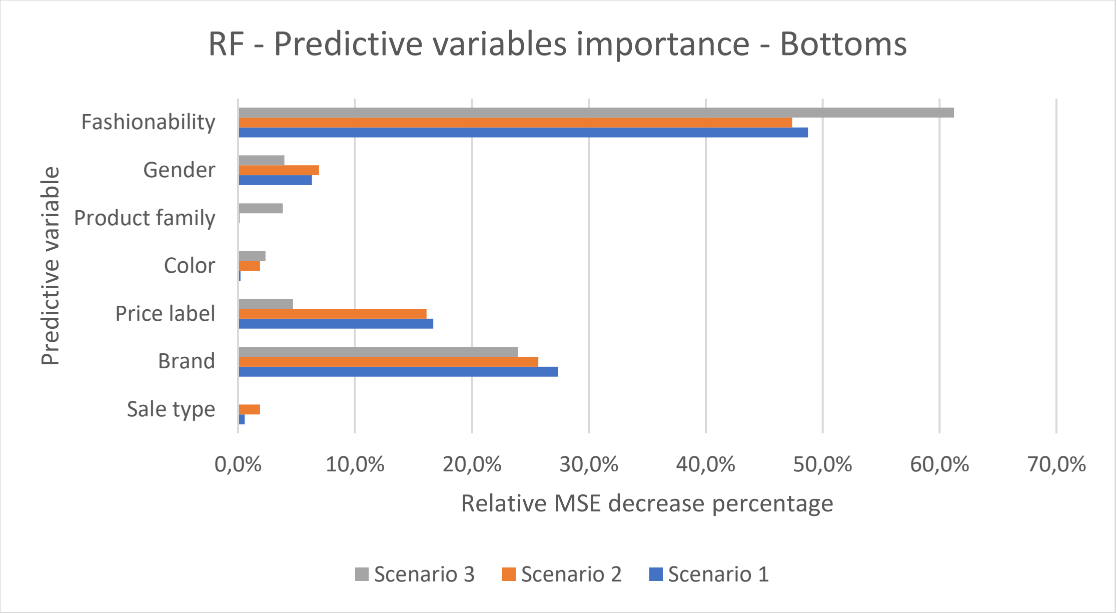


Figure 3 - Features importance results for bottoms

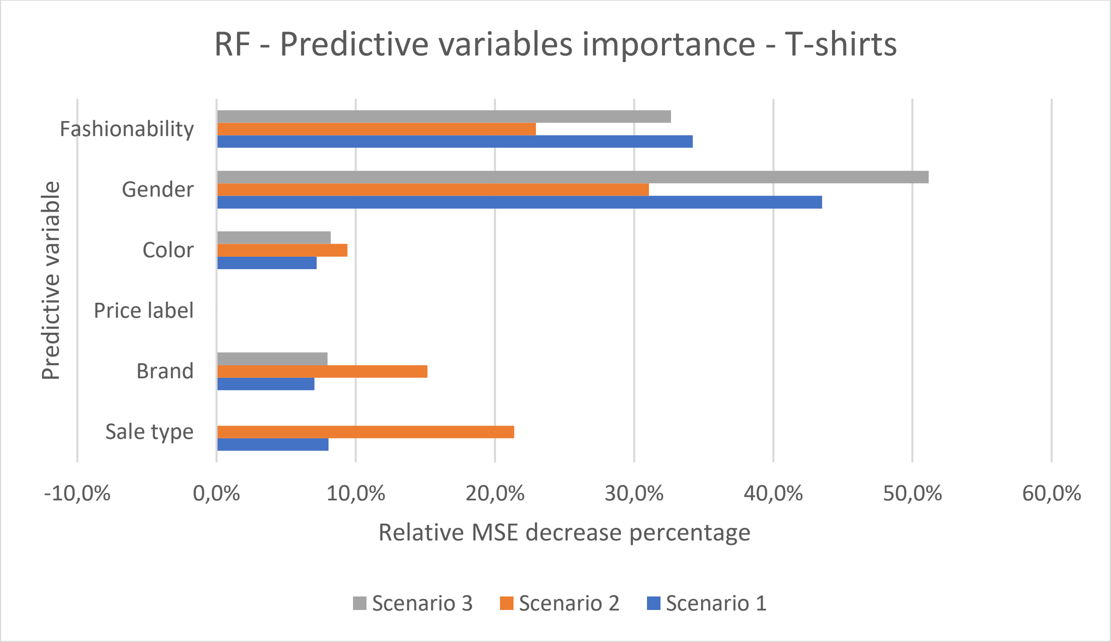


Figure 4 - Features importance results for T-shirts

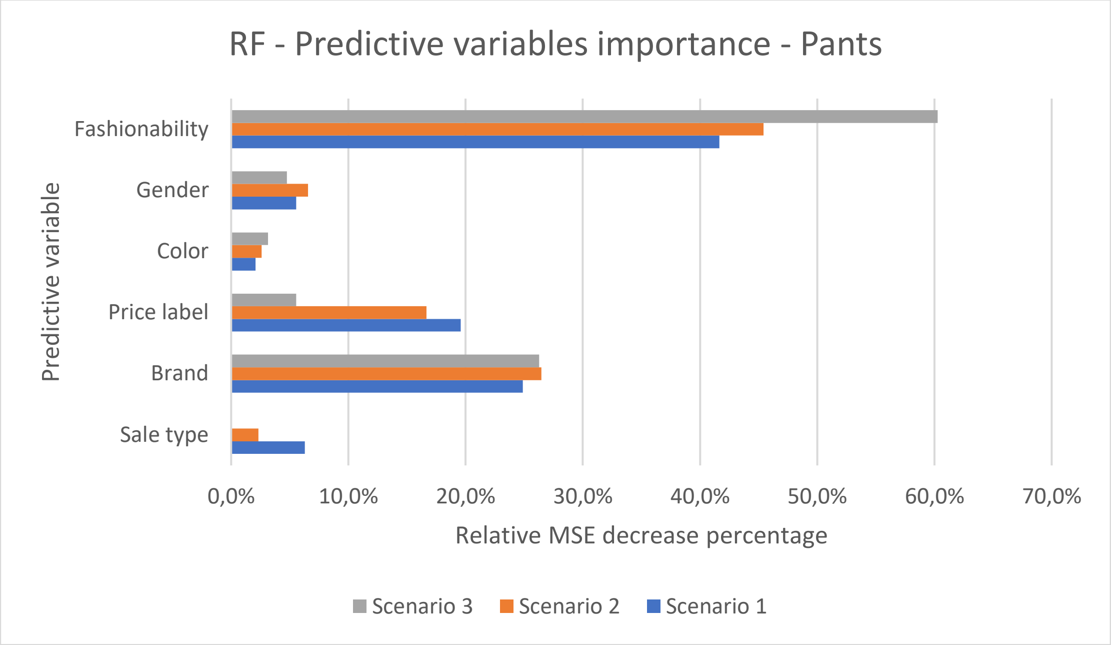


Figure 5 - Features importance results for Pants

ANEXO C: Catboost features importance remaining results

Considering:

- Scenario 1 – Units sold
- Scenario 2 – Units sold per week
- Scenario 3 – Ratio

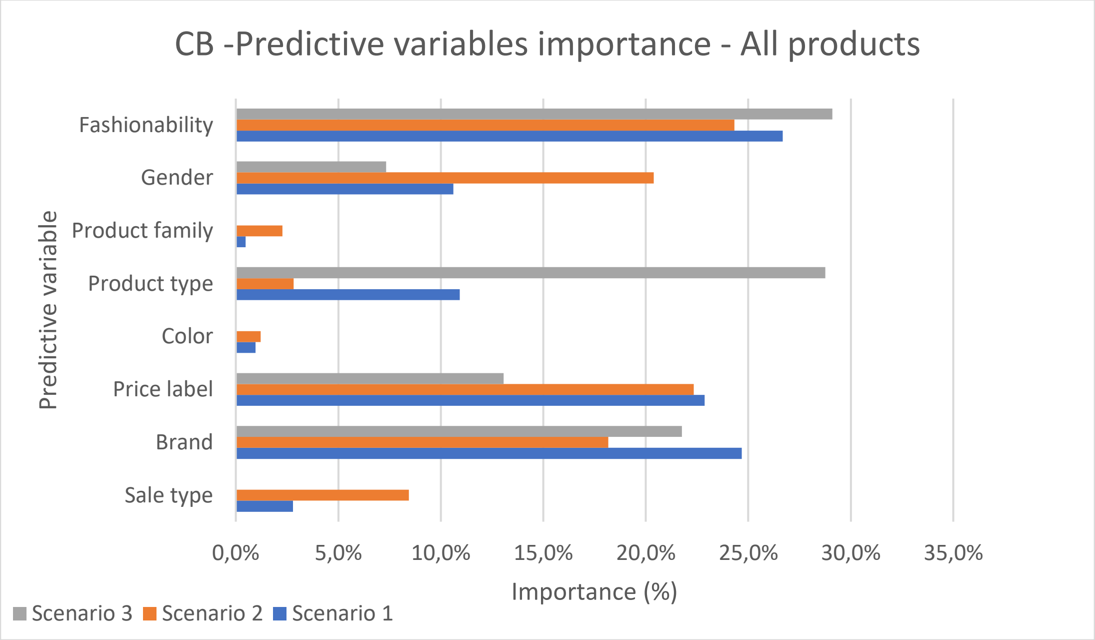


Figure 1 - Features importance results for all products

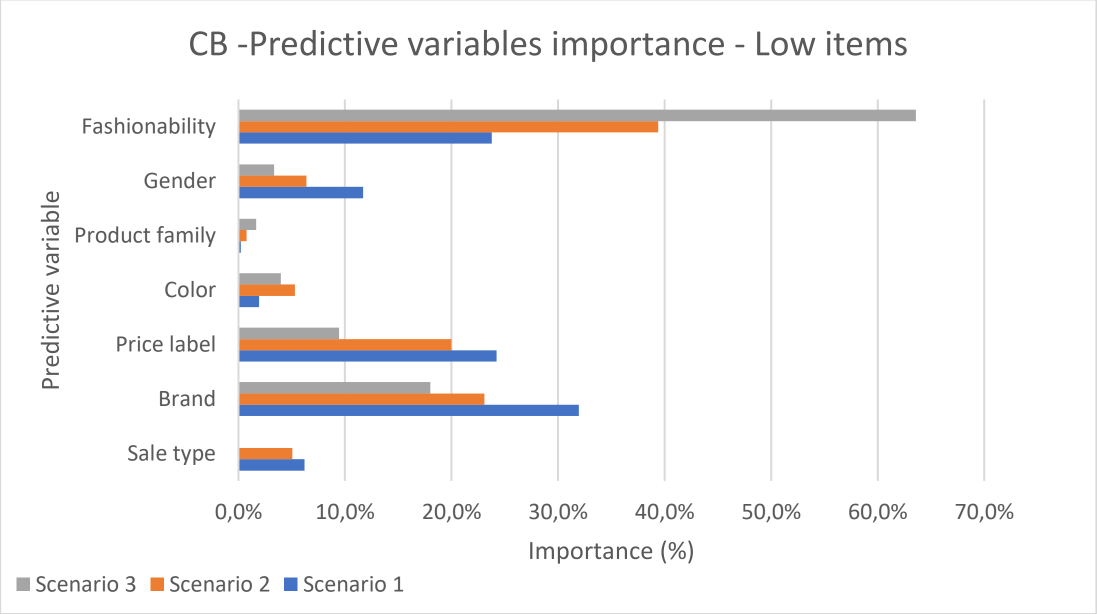


Figure 2 - Features importance results for low items

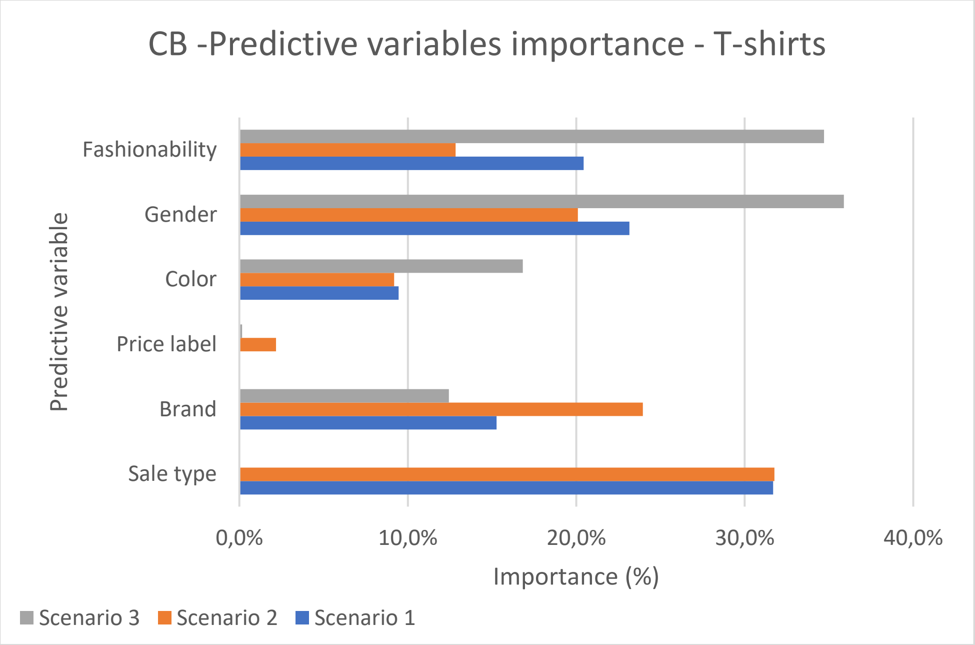


Figure 3 - Features importance results for T-shirts

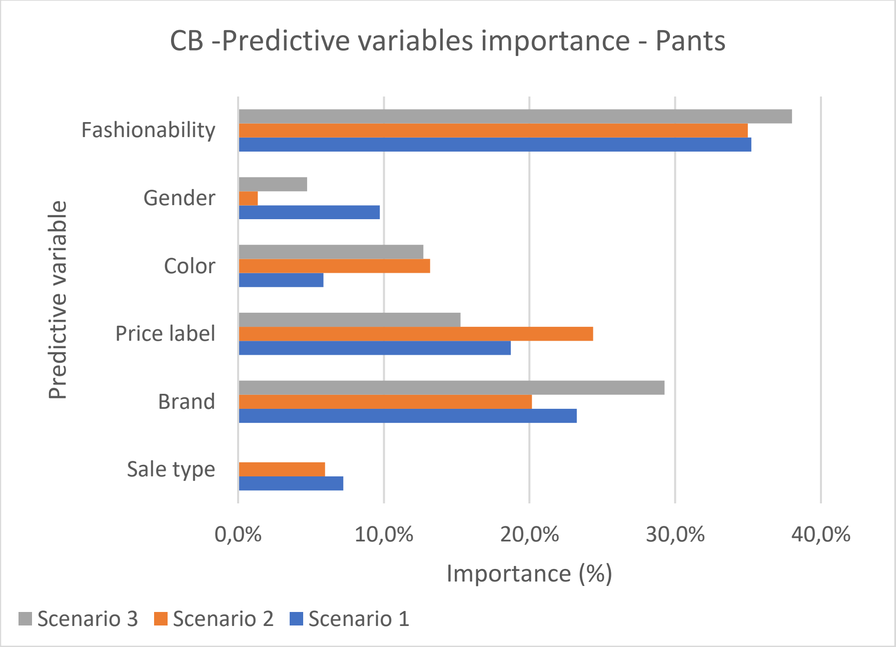


Figure 4 - Features importance results for pants