LU-TP 22-05
May 2022

Comparison of Machine Learning Algorithms in Predicting the Age Distribution Parameters of H&M Product Customers

Leyla Ismayil

Department of Economics, Lund University

Master thesis (DABN01, 15 ECTS) supervised by Krzysztof Podgórski

LUND UNIVERSITY

Abstract

Over the past decade the fashion industry has shifted towards dynamic product assortments with shorter life cycles. As a result, the role of the analysis of the product sales has increased and became crucial for fashion retailers. Using H&M Group's dataset on their product sales, I analyze and compare the performance of two machine learning algorithms in predicting the standard deviation and average age of product customers. These algorithms are Random Forest and Artificial Neural Networks. Since both dependent variables are estimated with noise, I fitted the models to the dataset with products having only a high number of observations. The paper describes in detail the performance of each of the algorithms and compares the accuracy. Random Forest works better in predicting the standard deviation of the age of customers per product, while ANN shows slightly better performance in predicting the average age of product customers. Both models perform better on a restricted sample and the performance of the models increases significantly while predicting the standard deviation of age.

Table of Contents

List of Figures

List of Tables

1. Introduction

In today's world, every company has a set of data that they collect over a certain period, maintain, and analyze. The analysis of the data is crucial for companies as by doing this they can improve and evaluate their future growth. Data analytics is implemented in various fields including the rapidly developing fashion industry.

The fashion industry has become more dynamic after the invention of the internet and e-commerce. The industry is now impacted by social media influencers, multichannel competition as well as fluctuations in the economy. All these factors have contributed to the reduction of the lifecycle of products as well as increased the variability in supply (Kharfan, 2020). As a result of these changes, many fashion retailers now work with dynamic assortments and products which have short life cycles. Hence, it is crucial for the fashion retailer companies to properly analyze sales of their products. In general, by better understanding sales of their products companies can improve their inventory management, increase profitability and optimize consumer targeting.

There are several characteristics of product sales understanding of which can generate valuable insights for the companies. For example, analysis can be performed to predict number of sold items or generated revenue of a given product considering the costs spent to make this product and other characteristics. The analysis can be further expanded if information on product customers is available. For example, demographic properties of the product audience can be studied. Several data-driven studies have been conducted on fashion products and most of them brought in computational models to make predictions (Kharfan, 2020).

In this study, I focus on explaining the age distribution of H&M product customers. In particular, I compare two machine learning algorithms in predicting the standard deviation and average age of product customers given the characteristics of products. These algorithms are Random Forest and Artificial Neural Network (ANN) which are commonly used in the literature. (see for example Alazwari A, Abdollahian M, Tafakori L, Johnstone A, Alshumrani RA, et al., 2022).

The dataset used in the study has been taken from the Kaggle.com website and covers the information about the products of Swedish retail company – H&M Group. The dataset covers transaction performed by customers between 2018 and 2019. With 100,010 unique products in my sample, my study contributes to the understanding of which machine learning algorithm

performs better in predicting the standard deviation and average age of product customers of H&M. The independent variables used in the study describe the characteristics of each product. These variables are product type, color, graphical appearance, price, number of product customers, the percentage of active members of H&M per each product. Since the dependent variables are estimated with noise, a restricted subsample of the dataset which contains only products purchased by more than 50 customers has also been created.

After thorough analysis of the dataset, Random Forest and ANN algorithms are used to obtain the predictions. The comparison is then made based on the RMSE and R-square estimators. The results of the study indicate that in predicting the average age of customers per product ANN model slightly outperforms the Random Forest, whereas in predicting the standard deviation the Random Forest model shows considerably better results. In addition, the standard deviation predictions become significantly better after fitting the model to the restricted sample.

There are several contributions of this study. First, the study provides recommendations for business analysts on which machine learning algorithm to use for a given parameter of age distribution. At the same time, the study highlights the importance of noise in estimation of distribution parameters. Second, the study provides guidelines for fashion retailers on how their data can be used to approximate the age group of their products. Using this information, these firms can improve their product recommendations, and as a result, increase the performance of their advertisement campaigns. Third, customers, in their turn, are potentially able to receive more relevant ads and recommendations from sales consultants, which would consider their age more systematically. Last but not least, government agencies and clothing donation organizations can use the findings of this research to distribute clothing to people in need in a more optimal way.

The rest of the paper is organized as follows. Section 2 reviews the contributions in literature. Section 3 gives a detailed description of the datasets used in the study and shows the summary statistics of the main variables used. Section 4 describes the methodology behind the estimation of the models used in the study and Section 5 shows the main results. The conclusion and the further investigations of the study are covered in Section 6.

2. Literature Review

In this section of the paper, I discuss existing literature related to this study. First, I mention business reports which explain the importance of customer age analysis for the companies. Second, I discuss academic articles which used Random Forest and Artificial Neural Networks algorithms to predict the mean or standard deviation of the variable.

2.1. The Importance of Customer Age Analysis for the Companies

The age of customers has been the point of interest for companies and this topic has been extensively researched over the past decades. A number of business reports explain the importance of analysis of the age of customers. It can be concluded that the age of customers is an important challenge for companies which also presents an incredible opportunity for growth. “The customer demographic attributes such as age and gender play a core role that may enable companies to enhance the offers of their services and target the right customer in the right time and place" (Mousa Al-Zuabi, Jafar and Aljoumaa, 2019). The article “Everyone is not your customer” published by Danny Shepherd discusses how the lack of demographic data leads businesses to lose their money. The article states that age is one of the most powerful metrics for targeting an ideal audience and explains the bid modifications the companies can do once they know the age category of buyers of their products. As an example, it shows leather jackets and suggests increasing the price for customers in the upper age group since these are the ones most likely to purchase the product. All these modifications serve to optimize daily expenses and increase ROI (Danny Shepherd, 2021). Another research also discusses the age of consumers and how it affects the desire for new products. The study indicates that there are age patterns in product-specific innovativeness and variety in product categories based on the results that the desire to switch from one product to another decreases with age.

2.2. Theoretical Framework

Several similar studies have been held using Random Forest and Artificial Neural Networks models. A number of papers describe the performance of these two algorithms and some of them focus on predicting the age variable. One of them is research held by Visit Limsombunchai, Mike Clemes and Amy Weng (2005) the aim of which is to identify which age group is the most popular among electronic banking users. The study compares the

predictive power of feed-forward neural networks and logistic regression in analyzing the age groups of consumers and their choices between electronic banking and non-electronic banking. According to the results, ANN outperformed the logistic model in predicting the age category of consumers. Another study held by Ghaida Riyad Mohammed, Jaffa Riad Abu Shbikah, Mohammed Majid Al-Zamili, Bassem S. Abu-Nasser, and Samy S. Abu-Naser (2020) uses ANN with a multi-layer model to predict the age of abalone from physical measurements. The results have shown 92% accuracy indicating that the algorithm is capable of predicting the age of abalone. Alazwari A, Abdollahian M, Tafakori L, Johnstone A, and Alshumrani RA (2022) also researched the best models to predict the age of people at the beginning stages of type 1 diabetes (T1D). The study has shown that out of the models fitted to the data Random Forest predicted the age at onset of T1D with the highest values of R2 and smallest values of RMSE.

Further studies have used ANN and RF to predict the mean and standard deviation of the certain variables. One of them is a study held by Ding (2018) the aim of which is to get accurate predictions for the weighted mean of temperature while the zenith wet delay is converted to precipitable water vapor. The ANN was validated and compared with four other models and the results show that ANN outperformed all four models in predicting the weighted mean of temperature.

From this body of literature, it can be concluded that Random Forest and Artificial Neural Networks are two of the most commonly used models which show better results in predicting the age related variables.

3. H&M Data Description

In this section, I discuss the datasets used in the study, the definition and construction of required variables, and provide a descriptive summary statistics of variables used in the analysis.

3.1. Sources of Data

As mentioned before, the dataset used in this research is from the Kaggle.com website. It has been published by H&M Group – a Swedish-based clothing retailer. In total three large datasets are cleaned and merged to be used in the model evaluation part of the research.

3.1.1. Articles Dataset

One of the datasets used in this study is the "Articles" dataset which describes each product of H&M Group purchased by the customer. The dataset consists of 105,542 observations. Each of the observations corresponds to one unique product, and variables describes the characteristics of the product. The dataset contains information about product type, product name, graphical appearance of the product, and the garment group to which the product belongs. For 105,542 unique products, there are 130 types of products, each of which belongs to one of 19 garment groups. There are 30 different graphical appearance indicators for each product which are "Solid", "Metallic", "Denim", etc. In addition, the dataset also contains information about the actual color of each product. Each of these unique colors has a corresponding perceived color which is defined by H&M Group based on how the colors are perceived by customers. There are 52 unique actual colors and 8 unique perceived colors. For example, the products with the actual colors "Black", "Dark-blue", "Dark-Grey", "Dark-Red", and "Brown" correspond to the perceived color "Dark" and colors "White", "Silver", "Light orange", "Light blue" correspond to perceived color "Light".

3.1.2. Customers Dataset

The other dataset used in the study is the "Customers" dataset which gives information about each customer of H&M Group who purchased the products from the "Articles" dataset. The dataset consists of 1,371,980 observations. Each observation corresponds to an individual customer who purchased an article from an H&M store. The dataset shows the status of membership of each customer indicating whether the customer is an "active" member of H&M. The main variable of interest is the "Age" variable which shows the corresponding age of each customer who made a purchase.

3.1.3. Transactions Dataset

Another available data is the "Transactions" dataset which contains information about transactions made by each customer from the "Customers" dataset to purchase the product from the "Article" dataset. In total, there are 31,788,324 observations in the dataset. Each observation is a transaction made by one customer to buy one product. Variables of the dataset describe the date of the purchase occurrence, the price paid for each product, and the sales channel of each purchase, meaning whether the purchase has been made online or offline. As

mentioned before, the transactions are made during the 2018 and 2019 years. For privacy reasons, the price variable is scaled by H&M Group. This does not affect results of this study since only the relative prices are important.

3.1.4. Articles Images

H&M Group also published the images of all the products from the "Articles" dataset. In total there are 105,542 images and all of them are in jpeg format. Due to time constraints, this dataset is not used in this research; however, it can be very important in predicting the age distribution parameters. For further investigation, deep learning algorithms can be used to analyze the images of all the products. The patterns of the products can be identified and clustered by using k-means clustering. It can lead to the creation of a new variable that will describe the patterns of each product, and the impact of this variable on the prediction of average age and standard deviation of age can further be investigated.

3.2. Definition and Construction of Variables

In this section, I explain the creation of the final dataset used for model fitting and define the list of dependent and independent variables used in the study.

My analysis is performed on the final dataset which is constructed by combining "Articles", "Customers" and "Transactions" datasets. Information across three datasets is aggregated to the product level. 5,532 observations with missing values on responsive variables have been removed from the final dataset as explained below. Hence the merged dataset is comprised of 9 variables (7 dependent and 2 independent) and 100,010 observations each of which corresponds to a single unique product.

Out of 9 variables, 4 are categorical and describe product type, graphical appearance, the actual color, and the perceived color of each product. For all these categorical variables the corresponding dummy variables have been created to be used in the analysis part of the paper. The other 5 variables are numerical and corresponds to the number of product customers, the percentage of customers who has an active membership in H&M, the average price of product, the average age of customers purchasing each product, and standard deviation of the age of customers purchasing each product. All the numerical variables have been calculated based on the information from the "Articles", "Customers" and "Transactions" datasets.

The mean and the standard deviation of the age of product customers are estimated with the noise like any other estimators. For this reason, I will also perform the analysis by restricting the dataset to have at least 50 customers per product. The dataset consisting of only products with more than 50 customers consists of 57,054 observations. From now on, I will refer to the final dataset containing all the products as the "full dataset" and to the final dataset containing only products with at least 50 customers as the "restricted dataset".

Table 1 summarizes all the variables used in the study. There are 2 dependent variables in my final dataset, and both of them are created using the age variable from the "Customers" dataset. The first variable is "avg_age" (μ_j) and corresponds to the average age of all the customers per product. More precisely, the variable has been calculated as follow:

$$\mu_j = \frac{\sum_{i=1}^{N} x_i^j}{N^j} \quad ,$$

$$where \;\; \mu_j - mean\ of\ age\ of\ customers\ of\ the\ j^{th}\ product$$
$$N^j - the\ number\ of\ customers\ of\ the\ j^{th}\ product$$
$$x_i^j - age\ of\ the\ i^{th}\ customer\ of\ the\ j^{th}\ produ\text{ct}.$$

The second variable is "std_age" (σ_j) which corresponds to the standard deviation of the age of customers who purchased each unique product. The variable is calculated based on the following formula:

$$\sigma_j = \sqrt{\frac{\sum(x_i^j - \mu_j)^2}{N^j - 1}} \quad ,$$

$$where \;\; \sigma_j - standard\ deviation\ of\ age\ of\ customers\ per\ each\ product$$
$$\mu_j - mean\ of\ age\ of\ customers\ of\ the\ j^{th}\ product$$
$$N^j - the\ number\ of\ customers\ of\ the\ j^{th}\ product$$
$$x_i^j - age\ of\ the\ i^{th}\ customer\ of\ the\ j^{th}\ produ\text{ct}.$$

Before calculating the mean and standard deviation of the age variable, some data cleaning is made. The redundant age observations (customers aged between 81-99) are replaced by missing values in order not to impact the mean and standard deviation values. Since most of these observations are the age of customers doing online purchases, these observations might

arise from the fact that while choosing their age on the website people have not identified their year of birth and just chose one of the first years appearing on the page. In addition, as mentioned earlier, for the standard deviation of age 5,532 missing values are observed since for some products there is only one customer who has purchased it. These observations are not used in estimations.

Table 1 Description of Variables used in the study

The table provides brief description of the variables used in this study. The variables are discussed in detail in Section 3.2.2.

Variable	Type	Description
product_type_name	categorical	type of the product Ex: top, sweater, trousers
graphical_appearance_name	categorical	the graphical appearance of the product, Ex: Solid, Metallic
color_group_name	categorical	color of the product
perceived_color_value_name	categorical	color of the product perceived by customer
number_of_customers	discrete	number of customers purchased the product
active_share	continuous	the percentage of customers who are the active members of H&M Group
price	continuous	average price paid by the customers for the product
avg_age	continuous	average age of customers purchased the product
std_age	continuous	standard deviation of age of customers purchased the product

There are 7 independent variables in the final dataset (see table 1). 4 out of 7 variables are qualitative and are product type, graphical appearance, perceived color, and actual color – taken from the "Articles" dataset. The corresponding dummy are defined and assigned for all these categorical variables by H&M. In total, there are 130 unique product types, 30 unique graphical appearance values, 52 unique actual colors, and 8 unique perceived colors in the dataset. The variables "number_of_customers" and "active_share" are calculated based on the information taken from the "Customers" dataset. The variable "number_of_customers" shows the total number of customers who purchased the particular product. The variable "active_share" shows for each product the percentage of customers who are active members

of the H&M Group. For the price variable, the average of all the prices paid by customers is calculated for each product.

All these variables are used to understand how accurately two machine learning algorithms can predict the standard deviation and average age of product customers.

3.3. Summary Statistics

In this section, I provide the summary statistics of the variables in the final dataset. This is done for the full and for the restricted samples.

Table 1 summarizes the quantitative variables used in this study for the full and restricted datasets. In general, if we look at the standard deviation of the age, we can see that the mean of this variable in both full and restricted samples does not differ too much being 11.1 and 12 respectively. The standard deviation of this variable differs a lot between the two samples. For the restricted sample, it is only 1.9, whereas for the full dataset, it is 3.2 which means that the more customers you require per product less noise there is and correspondingly, the lower is the standard deviation of the estimate. In addition, 25% of H&M products have a standard deviation of less than 9.6 which is less than the mean of this variable; whereas, the 75th percentile is 13.1 indicating that 25% of the products have a standard deviation of more than 13.1. If we look at the products that have more than 50 customers it is clear that the 25th percentile is higher being 11, while the 75th percentile does not change much. The fact that the 25th percentile is lower for all the products is because there are 357 products in the dataset which have a standard deviation equal to 0. These products have been mostly purchased by only 2 or 3 customers. In the restricted dataset, the minimum value for the standard deviation is 2.91.

The mean of average age variable is around 37 years, and it does not change much with the restriction on number of customer while the standard deviations of this estimate falls slightly from 4.8 to 4.1.

Table 2 Summary of quantitative variables

The table provides summary statistics for quantitative variables used in this study. The table reports mean, standard deviation, 25th percentile, median, 75th percentile and number of observations (N) for each of the variables. The top panel reports the statistics for the full dataset, the bottom panel reports the statistics for the restricted dataset.

Full dataset	Mean	St. Dev.	Pctl25	Median	Pctl75	N
number_of_customers	*317806*	*806319*	*17*	*73*	*306*	*100010*
active_share	*0.971*	*0.055*	*0.968*	*0.988*	*1*	*100010*
price	*0.029*	*0.026*	*0.015*	*0.023*	*0.034*	*100010*
avg_age	*37.5*	*4.8*	*34*	*37.075*	*40.6*	*100010*
std_age	*11.1*	*3.2*	*9.6*	*11.7*	*13.1*	*100010*

Restricted dataset	Mean	St. Dev.	Pctl25	Median	Pctl75	N
number_of_customers	*544037*	*1010130*	*111*	*248*	*597*	*57054*
active_share	*0.976*	*0.027*	*0.969*	*0.983*	*0.992*	*57054*
price	*0.031*	*0.023*	*0.016*	*0.025*	*0.039*	*57054*
avg_age	*36.9*	*4.1*	*33.7*	*36.4*	*39.7*	*57054*
std_age	*12*	*1.9*	*11.1*	*12.2*	*13.2*	*57054*

Table 3 summarizes the number of products, the number of customers, the mean average age, and the mean the standard deviation of the age for each graphical appearance of the products of H&M. In general, it is clear that most of the products of H&M have the "solid" graphical appearance (47,236 products). All these "solid" graphical appearance products have been purchased by 17,847,926 customers, which is the highest number in the corresponding column. The average age is mostly between 36-38, and the highest mean average age is 41 which corresponds to the customers purchasing products with an “Argyle” graphical appearance. The products, graphical appearance of which is “transparent”, have the highest standard deviation of 12.6, and this means these products have been purchased by a wide range of age groups.

Table 4 summarizes all the above-mentioned indicators for each of the product groups. If we look at different groups of products, we can see that the mean average age is higher for nightwear, cosmetics, and garment and shoe care. The standard deviation of customers' age for different product groups is mostly between 10-11. The lowest standard deviation is for interior textile being only 7.6, which means that customers purchasing the products of this group are from the same age groups, although this indicator is a bit noisy since only 8 customers purchased products of this group. Products with high standard deviation of age are bags, furniture, and stationery.

Summary statistics for product type, actual color and perceived color can be found in Appendix.

Table 3 Summary statistics for graphical appearance variable

The table provides summary statistics for graphical appearance variable used in this study. The table reports the number of observations (N), total number of customers, the mean of average age and the mean of standard deviation of age for all 30 graphical appearance names in the dataset.

graphical_appearance_name	N	Sum of customers	Mean of average age	Mean of st. dev. of age
Solid	*47236*	*17847926*	*37.1*	*11.3*
Stripe	*4768*	*1450561*	*37.3*	*10.9*
All over pattern	*16214*	*3972106*	*37.6*	*10.9*
Melange	*5619*	*1900407*	*38.6*	*11.6*
Transparent	*81*	*18068*	*36.1*	*12.6*
Metallic	*319*	*61434*	*36.2*	*11.4*
Application/3D	*1257*	*137579*	*37.8*	*10.6*
Denim	*4694*	*1970624*	*38.3*	*10.8*
Colour blocking	*1729*	*240101*	*38.5*	*10.1*
Dot	*655*	*243903*	*36.8*	*11.1*
Other structure	*1433*	*726803*	*37.0*	*11.3*
Contrast	*350*	*174773*	*36.5*	*11.7*
Treatment	*552*	*113090*	*39.5*	*10.3*
Check	*2068*	*591307*	*38.3*	*12.4*
Chambray	*309*	*35688*	*38.6*	*9.9*
Front print	*3017*	*321014*	*39.5*	*10.0*
Glittering/Metallic	*895*	*187344*	*36.4*	*11.0*
Mixed solid/pattern	*992*	*71840*	*39.2*	*9.3*
Placement print	*2861*	*431510*	*38.0*	*10.7*
Other pattern	*475*	*40236*	*38.7*	*9.9*
Neps	*63*	*30301*	*38.4*	*11.7*

Embroidery	*1117*	*311279*	*35.8*	*11.7*
Jacquard	*785*	*174782*	*38.0*	*11.7*
Unknown	*49*	*9019*	*37.6*	*12.6*
Lace	*1469*	*589684*	*35.2*	*11.2*
Argyle	*14*	*1101*	*41.0*	*10.6*
Slub	*142*	*21131*	*38.5*	*10.6*
Mesh	*82*	*40718*	*34.5*	*11.2*
Sequin	*758*	*69243*	*39.6*	*10.4*
Hologram	*7*	*164*	*34.8*	*10.9*

Table 4 Summary statistics for each product group

The table provides summary statistics for product group name variable used in this study. The table reports the number of observations (N), total number of customers, the mean of average age and the mean of standard deviation of age for all 19 product group names in the dataset.

product_group_name	Number of observations	Sum of customers	Mean of average age	Mean of st. dev. of age
Garment Upper body	40395	12550814	37.9	11.4
Underwear	*5234*	*2565644*	*36.0*	*10.9*
Socks & Tights	*2289*	*685578*	*37.3*	*10.3*
Garment Lower body	*18828*	*7045208*	*38.4*	*10.8*
Accessories	*10441*	*1599006*	*36.1*	*11.5*
Items	*15*	*5427*	*34.4*	*12.3*
Nightwear	*1804*	*348101*	*38.8*	*10.3*
Unknown	*113*	*97040*	*36.9*	*12.2*
Underwear/nightwear	*47*	*553*	*36.5*	*9.6*
Shoes	*5006*	*745299*	*37.9*	*10.8*
Swimwear	*3047*	*2579140*	*35.2*	*10.5*
Garment Full body	*12691*	*3551998*	*36.7*	*11.0*
Cosmetic	*44*	*1496*	*39.5*	*10.5*
Interior textile	*3*	*74*	*35.1*	*7.7*
Bags	*25*	*7313*	*35.0*	*12.7*
Furniture	*13*	*533*	*37.2*	*12.2*
Garment and Shoe care	*9*	*279*	*40.5*	*11.7*
Stationery	*5*	*229*	*32.3*	*12.3*
Fun	*1*	*4*	*32.0*	*9.8*

Figures 1 shows the distribution of the average age of customers and the standard deviation of the age of customers for all the products of H&M. Most of the products in the dataset have been purchased by adults who are in average between 30-45 years old. This can be seen from

the histogram (Figure 1). In total 3,183 products have been purchased by customers average age of which is less than 30 years and 6,229 products have customers with an average age of more than 45. Popular products purchased on average by young customers (less than 30 years old) are dresses, tops, and t-shirts. There are only 53 products which have an average age of customers of more than 60. The most popular products with such average age are jackets and sweaters.

If we look at the standard deviation of the customers' age per product, we can see that for most of the products the standard deviation is 10-12. There are 5,610 products with a standard deviation of less than 5 which means that these products are mostly purchased by customers of one age category. The most popular products with such standard deviation are shorts, t-shirts, and tops. 4,899 products have a standard deviation of more than 15, and hence are purchased by teenagers as well as adults and elder people. The popular products of such standard deviation are sweaters, jackets, and dresses.

Figure 1 Frequency of the standard deviation and average age of product customers

The figure shows the distribution of average and standard deviation of age of product customers in the full dataset.

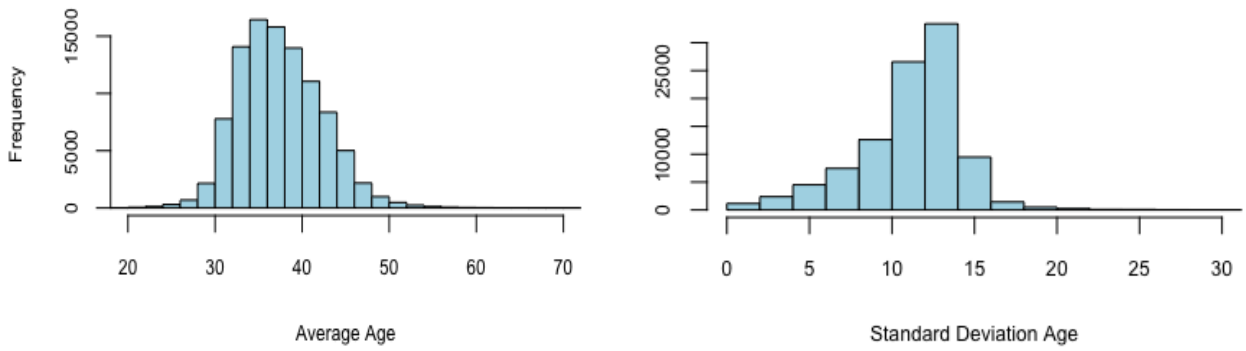
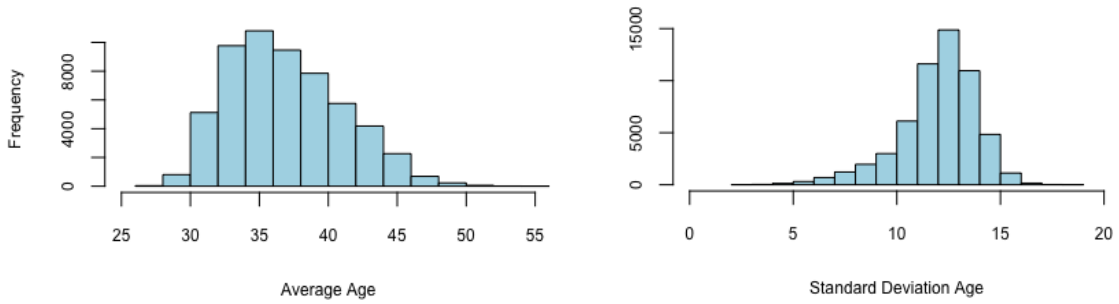


Figure 2 shows the distribution of the average age of customers and the standard deviation of age of customers for the products of H&M in the restricted dataset. These products have been purchased mostly by adults who are, on average, between 30-40 years old. If we compare it with the histogram for all the products, we can see that the average age range for products with a high number of customers is between 25-55, whereas for all the products, the range is between 20-70. There are only 833 products the average age of customers of which is less than 30, and 1,867 products the average age of customers of which is more than 45 years. If we look

at the standard deviation of the customer ages per product, we can see that for most of the products, the standard deviation of customers' age is 10-15. There are only 161 products with a standard deviation of less than 5, showing that these products are mostly purchased by customers of the same age group. There are only 1,309 products that have a standard deviation of more than 15, which means that they are purchased by the customers of different age categories.

Figure 2 Frequency of the standard deviation and average age of product customers in the restricted sample

The figure shows the distribution of average and standard deviation of age of product customers in the restricted dataset



4. Methodology

In this section, I discuss two models fitted to the datasets. I give a general overview of the algorithms and explain the tuning of hyperparameters of the models.

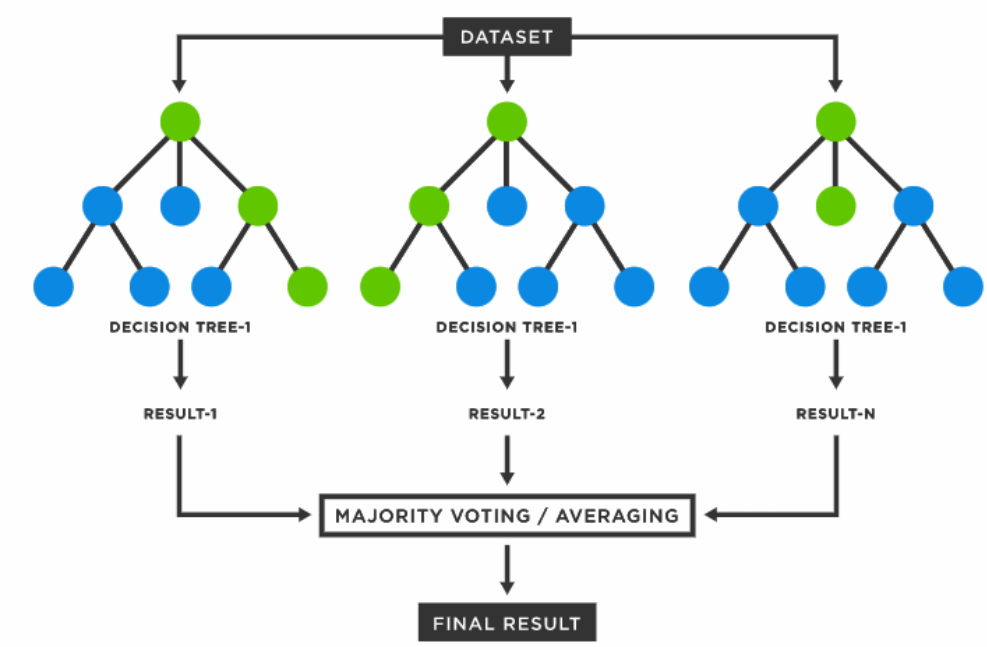
4.1. Random Forest General Understanding

Random Forest is a supervised machine learning algorithm used for regression as well as for classification problems. The algorithm is highly adaptable to a wide range of datasets due to its simplicity and flexibility. It is called forest because it builds a forest of decision trees by fitting them on different bootstrapped samples i.e. trees are trained on different datasets (Mbaabu, 2020). Each decision tree is built by using a random subset of variables at every split. All these protect trees from errors and incorrect predictions. Selected random bootstrapped dataset usually uses approximately two-thirds of the data. After creating the trees, the algorithm runs each observation on each of the created trees and then takes the average of

the results of trees to improve the predictive accuracy and control the overfitting. By this, the algorithm assures more accurate results. Figure 4 shows the overview of a Random Forest algorithm.

Figure 3 Overview of Random Forest Algorithm

The figure shows the overview of the Random Forest model with N number of decision trees. The algorithm shows that the final results is obtained by averaging the results from N decision trees in case of regression problem, and by taking the majority voting in case of classification problem.



Source TIBCO Software Inc.

In general, the main advantage of using a random forest algorithm in this study is that it is very flexible and is incredibly efficient with all types of data, and there is a very small risk of overfitting, as long as the number of trees is enough, and the speed of getting predictions is faster since it uses only a subset of features. However, it should also be noted that the Random Forest algorithm is known as a "black box" since it is difficult to understand how and why a certain decision is made.

4.1.1. Hyperparameters Tuning

After cleaning the data and creating the final dataset, which is discussed in detail in section 3, I started to fit the Random Forest model by using the Skicit-Learn library. I use 80% of the dataset to train and 20% of the data to test the model. Then I choose the corresponding hyperparameters to train the model. In Random Forest, the hyperparameters are the number of

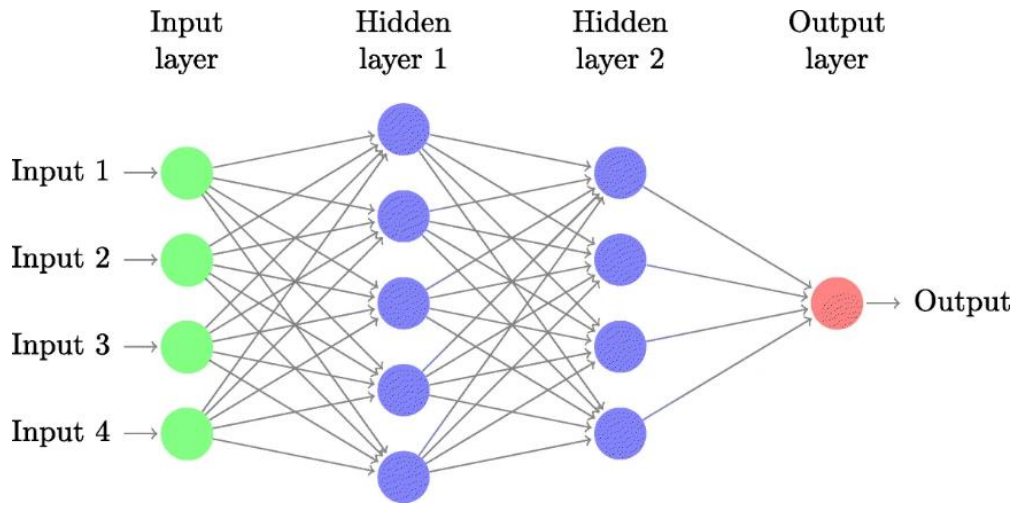
decision trees in the forest and the number of variables used at every split of the tree. I adjust the hyperparameters to improve the model performance by defining a grid of values and evaluating the range of these values for each hyperparameter. With each combination of values, 5-fold cross validation has been performed, and by this, random samples have been chosen from the grid. As a result, I get the best parameters chosen for the number of trees and the maximum depth of each tree which are 500 and 12 respectively.

4.2. Artificial Neural Networks General Understanding

The second model fitted to the full and restricted datasets is Artificial Neural Networks. ANNs are nonparametric modeling tools that have the potential to perform complex function mapping with the desired accuracy. The models are very flexible and are composed of several layers each having a different number of nodes. Each node receives output from the other nodes and then processes it through activation function and transforms the result to other nodes or the final result. Figure 4 shows an ANN architecture with 1 input layer, 2 hidden layers, and 1 output layer.

Figure 4 Artificial Neural Networks Architecture

The figure shows the architecture of the ANN model with 1 input layer, 2 hidden layers and 1 output layer. The input layer has 4 inputs, the 1st hidden layer has 5 hidden neurons, the 2nd hidden layer has 4 hidden neurons. The model predicts 1 output.



Source BioMed Central Ltd

4.2.1. Summary of ANN Model

This is well known that ANNs can be varied in an infinite number of ways. Several layouts are tested with a different number of hidden layers and hidden neurons to confirm that the chosen network layout is good enough. I have decided to test 3 configurations. In the first configuration, my network has 3 hidden layers with 64, 16, and 4 hidden neurons in each layer respectively. In the second configuration, the model has 2 hidden layers with 36 and 6 hidden neurons in each layer respectively. In the third configuration, the model has a single hidden layer with 15 hidden neurons. All 3 networks have been trained on the training data 30 times. During each training, 25% of the data has been considered validation data, and the network which performed best on the validation data was saved.

The activation function in all of the networks is chosen to be Rectified Linear Unit (ReLU):

$$f(x) = \max(0, x)$$

5. Results

In this section, I discuss the main results of the paper. As mentioned before, the main aim of this study is to compare the performance of the two machine learning algorithms in fitting the standard deviation and average age of product customers. These algorithms are Random Forest and Artificial Neural Networks (ANN). Since both of the dependent variables are estimated with noise, the analysis has been performed on both full and the restricted samples. The measures of accuracy which is used to compare the models are RMSE and R-square. First, I describe the performance of the Random Forest model in predicting the average and the standard deviation of customers' age. Second, I describe the performance and predictive power of the ANN model. Then I compare the findings from both algorithms and conclude the study.

5.1. Random Forest

First, I trained the model on the full dataset and then on the restricted one. At each iteration, my model generated 500 decision trees which have a maximum depth of 12 nodes. By fitting the model to the test set, I get the predictions and estimated the accuracy of the model. I looked at four accuracy measures to evaluate the performance of the model. Table 5 summarizes four

accuracy measures (MSE, RMSE, MAE, R-square) gained from fitting the model to two samples and predicting the average and standard deviation of the age of customers per product.

Table 5 The Results based on the random forest model

The table provides the summary for the results from fitting the Random Forest model to the datasets. The top table represents the results for full dataset, the bottom table for the restricted dataset. The accuracy measures shown are MSE, RMSE, MAE and R-square. All of the accuracy measures have been calculated based on the prediction of average age and standard deviation of age of product customers.

	Avg_Age_Full	Avg_Age_Restricted	Std_Age_Full	Std_Age_Restricted
MSE	19.4	13.6	7.96	2.39
RMSE	4.4	3.7	2.82	1.54
MAE	3.43	3	1.94	1.18
R-square	0.16	0.18	0.21	0.31

In general, it can be seen that the model performs better in predicting the standard deviation of age of customers rather than in predicting the average age. For both variables, the predictive power of the model improves once we restrict the dataset. The best results are obtained by fitting the Random Forest to the restricted dataset to predict the standard deviation of age of customers per product. According to Table 5, 31% of the variability in the standard deviation of the age of customers can be explained by the model which is considered to be substantial for the real-world data. For the full dataset, the R-square is lower (21%). For the average age of customers, the model predicts the data relatively less accurately. Only 16% of the variability in the average age of customers per product is explained by the model. When we limit the dataset to only products with more than 50 customers, the R-square estimator becomes slightly higher. If we look at RMSE and MAE, we see that the estimators for both variables also improve once we restrict the dataset. For the average age of customers, RMSE decreases by 16%, whereas for the standard deviation of the age of customers, it decreases by almost 50%. This indicates that the standard deviation of age is estimated with more noise than the average age of customers. From Table 2, we know that the mean average age of customers is 37.4 for the full dataset and 36.9 for the restricted dataset. Thus, RMSE for the average age of customers is 4,4 in the full dataset which means that RMSE is about 11% as large as the mean of the outcome. For the restricted dataset, the value is about 10%. The mean standard deviation of age is 11 and 12 for the full and restricted datasets respectively. For the standard deviation, the model predicts much better in the restricted dataset. If we look at RMSE for the standard

deviation of age in the full dataset, it is 2,82 which means that RMSE is about 25% as large as the mean of the standard deviation of age. However, in the restricted dataset RMSE is about 12% as large as the mean of the outcome.

5.2. ANN

As mentioned before, I created three models based on 3 configurations and fitted each of them to the full and restricted samples. Table 6 shows the accuracy results that I get by fitting these models.

Table 6 Result from fitting ANN to the full dataset

The table provides the summary for the results from fitting ANN model to the datasets. The top table represents the results for full dataset, the bottom table for the restricted dataset. The first two rows describe number of hidden layers and the number of hidden neurons in the model. The accuracy measures shown are MSE, RMSE, MAE and R-square. All of the accuracy measures have been calculated based on the prediction of average age and standard deviation of age of product customers. The best results obtained are shown in bold.

Full dataset	Avg_Age			Std_Age		
Hidden Layers	3	2	**1**	**3**	2	1
Hidden Neurons	[64;16;6]	[36;6]	**[15]**	**[64;16;6]**	[36;6]	[15]
MSE	20.82	19.51	**19.12**	**8.18**	8.69	8.3
RMSE	4.56	4.41	**4.37**	**2.86**	2.94	2.88
MAE	3.40	3.40	**3.38**	**1.97**	1.98	2.00
R-square	0.15	0.15	**0.16**	**0.19**	0.14	0.18

Restricted dataset	Avg_Age			Std_Age		
Hidden Layers	3	**2**	1	**3**	2	1
Hidden Neurons	[64;16;6]	**[36;6]**	[15]	**[64;16;6]**	[36;6]	[15]
MSE	13.30	**12.85**	14.59	**2.69**	2.89	2.70
RMSE	3.64	**3.58**	3.81	**1.64**	1.70	1.64
MAE	2.91	**2.87**	3.11	**1.25**	1.31	1.24
R-square	0.20	**0.23**	0.13	**0.23**	0.17	0.22

In general. the model performed better in a restricted dataset predicting both average and standard deviation of the age of customers. According to table 6, the model performs the best in predicting the average age with configuration 1 for the full sample and with configuration 2

for the restricted sample. Once we restrict the dataset, the R-square improves by 7 percentage points. The improved result for R-square is 23% which means that the model is explaining 23% of the variability in the average age of customers. MAE values also show an improvement after restricting the dataset and decrease by 15%. The improved value shows that the average distance of predicted values for the average age of customers from the actual values is 2,87. There is also an 18% improvement in RMSE estimators.

In predicting the standard deviation of age of product customers, the model built in configuration 3 gives the best result in both samples. In the restricted sample, the model explains 23% of the variability in the standard deviation of age of customers, which is 4 percentage points higher than in the full dataset. There is a significant improvement in RMSE values, which are 2,86 and 1,64 for the full and restricted samples respectively. As one can see, once we restrict the dataset, the RMSE is improved by almost 50%. MAE values are 1,97 and 1,25 respectively, showing an improvement of 37% after restricting the dataset.

5.3. Comparison of Two Models

In this section, I compare the results of Random Forest and ANN models, which are fitted to the full and restricted samples. For the comparison I focus on RMSE and R-squared measures. Table 7 illustrates the main results of the study.

Table 7 The results of two models

The table provides the RMSE and R-square measures of the two algorithms used in the study. The top table shows the results for Random Forest algorithm in both full and restricted samples. The bottom table shows the results for ANN algorithm in both full and restricted samples. Both algorithms predict the average age and standard deviation of age of product customers.

Random Forest	Avg_Age_Full	Avg_Age_Restricted	Std_Age_Full	Std_Age_Restricted
RMSE	4.40	3.70	2.82	1.54
R-square	0.16	0.18	0.21	0.31
Neural Network	Avg_Age_Full	Avg_Age_Restricted	Std_Age_Full	Std_Age_Restricted
RMSE	4.37	3.58	2.86	1.63
R-square	0.16	0.23	0.19	0.23

As mentioned before, both the Random Forest model and ANN improve their performances once the dataset is restricted to products with at least 50 customers. The standard deviation variable appears to be estimated with more noise than average age since the results significantly get better after looking only at the restricted dataset.

If we compare both models based on average age prediction accuracy, we can see that ANN did a slightly better job in predicting the outcome. Both of the models in the full dataset explained 16% of the variability in the average age of customers. However, when the dataset is restricted, the R-square estimator for the random forest is 18%, whereas for the ANN it increases by 5 percentage points. Thus, ANN works better in explaining the variability in the average age of customers of the products in the restricted dataset. RMSE values for the ANN in both datasets also show a slight improvement while predicting the average age. In the full dataset, the ANN's RMSE value is improving by 1%, whereas for the restricted dataset the value improves by 4%.

In predicting the standard deviation of customers' age, however, the Random Forest algorithm does a better job. In both datasets, the accuracy values are higher for the Random Forest model. In the full dataset, the Random Forest explains 21% of the variability in the standard deviation which is 2 percentage points higher than does ANN model. In the restricted dataset, the difference is more apparent. The Random Forest model explains 31% of the variability in the standard deviation, whereas ANN explains 23%. RMSE values also show an improvement of 2% and 6% in the full and restricted datasets respectively.

6. Conclusion

Motivated by the fact that fashion retailers now experience more dynamic assortments and products with a short life cycle, I study the relationship between different product characteristics and the age of customers purchasing the product. I look at the performance of two machine learning algorithms in predicting the average age and standard deviation of age of product customers. I use the rich dataset of Swedish retailer H&M which gives information on 100,010 products. In addition, since my outcome variables are estimated with the noise, I also fit my models to the restricted dataset which consists of only the products with more than 50 customers. The restricted dataset used in the study consists of 57,054 products. The

variables used in the study are product type, graphical appearance, actual color, perceived color, share of active members and the average price of the product as well as the number of customers per each product. I fit the Random Forest and Artificial Neural Networks algorithms to predict the average age and standard deviation of age. The methodology behind the comparison is to look at the RMSE and R-square estimators.

There are several findings from this study. First, in predicting the average age of customers the ANN model slightly outperforms the Random Forest. However, in predicting the standard deviation of the age of customers the Random Forest works considerably better and predicts the outcome with a higher accuracy. Second, both models improve their performance on the restricted dataset. The standard deviation predictions become significantly better once I restrict the dataset which indicates that the variable is estimated with higher noise than the average age.

Despite contributions of this paper, there are several limitations of this study. First, the models do not say anything about the skewness of the age distribution which might be of interest for fashion retailers. Second, the study has been conducted based on the dataset provided by H&M Group and the results do not necessarily hold for audiences outside of H&M. Hence, it would be interesting to know if these findings are robust to other retailers' customers. Third, some of the independent variables used in the study might be difficult to obtain. For example, the number of customers per each product is hard to identify when introducing new product. However, this variable itself, can in turn be estimated by using machine learning algorithms. Finally, I recognize the fact that there are potential omitted variables such as product quality or shelving which are potentially important in predicting the distribution parameters of age.

Further studies can utilize images dataset of products published by H&M. First of all, the images of all the products can be analyzed separately by using deep learning algorithms for image recognition. By this, certain patterns on the clothes can be detected, and, by using a k-means clustering algorithm, they can be grouped into certain clusters. A new "pattern" variable can be added to the dataset, and the impact of this variable on the outcome can further be studied by fitting the models. In addition, the purchase dates of the products can be used in the

analysis and the products can be separated by four seasons or special dates. By this, the predictions of the outcome can further be investigated based on seasonality or certain holidays.

References

DannyShepherd. (2021, October 28). *Target audience age groups & gender: Targeting guide*. Titan Growth. Retrieved May 10, 2022, from https://www.titangrowth.com/blog/everyone-is-not-your-customer-why-age-gender-targeting-matter/

Ahlfeldt, J. (2018, March 8). *The age of the customer: What it means and why you should care*. Retrieved May 10, 2022, from https://www.linkedin.com/pulse/age-customer-what-means-why-you-should-care-julia-martin/

Al-Zuabi, I.M., Jafar, A. & Aljoumaa, K. Predicting customer's gender and age depending on mobile phone data. *J Big Data* **6,** 18 (2019). https://doi.org/10.1186/s40537-019-0180-9

Alazwari A, Abdollahian M, Tafakori L, Johnstone A, Alshumrani RA, et al. (2022) Predicting age at onset of type 1 diabetes in children using regression, artificial neural network and Random Forest: A case study in Saudi Arabia. PLOS ONE 17(2): e0264118. https://doi.org/10.1371/journal.pone.0264118

Ding, M. A neural network model for predicting weighted mean temperature. *J Geod* **92,** 1187–1198 (2018). https://doi.org/10.1007/s00190-018-1114-6

Mbaabu, Onesmus. "Introduction to Random Forest in Machine Learning." *Engineering Education (EngEd) Program | Section*, 20 Dec. 2020, www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

Koehrsen, Will. "Improving the Random Forest in Python Part 1." *Medium*, 8 Jan. 2018, towardsdatascience.com/improving-random-forest-in-python-part-1-893916666cd.

Gan, Christopher, et al. *CONSUMER CHOICE PREDICTION: ARTIFICIAL NEURAL NETWORKS versus LOGISTIC MODELS*. July 2005.

Loureiroa, A.L.D., et al. *Exploring the Use of Deep Neural Networks for Sales Forecasting in Fashion T Retail*. Aug. 2018.

Probst, Philipp, et al. *Hyperparameters and Tuning Strategies for Random Forest*. Feb. 2018.

Jafar, Assef, et al. *Predicting Customer's Gender and Age Depending on Mobile Phone Data Open Access*. 2019.

V. Rodriguez-Galiano, et al. *Machine Learning Predictive Models for Mineral Prospectivity: An Evaluation of Neural Networks, Random Forest, Regression Trees and Support Vector Machines*. 2015.

Chen,I.-F.;Lu,C.-J. Demand Forecasting for Multichannel Fashion Retailers by Integrating Clustering and Machine Learning Algorithms. Processes 2021, 9, 1578. https://doi.org/10.3390/pr9091578

Kharfan, Majd, et al. *A Data-Driven Forecasting Approach for Newly Launched Seasonal Products by Leveraging Machine-Learning Approaches*. 23 June 2020.

Appendix

Table 8 The summary statistics for each product type

The table provides summary statistics for product type variable used in this study. The table reports the number of observations (N), total number of customers, the mean of average age and the mean of standard deviation of age for all 130 product types in the dataset.

Product Types	N	Customers	Mean of Avg. Age	Mean of St. Dev.
Trousers	10608	4216559	39,0	11,0
Dress	9950	3238112	37,0	11,3
Sweater	8774	2782870	38,1	11,6
T-shirt	7475	2203350	37,6	10,8
Top	3957	1583258	37,2	11,2
Blouse	3803	1504728	37,6	12,0
Vest top	2860	1413974	36,3	10,7
Bra	2131	1335167	34,2	11,1
Shorts	3765	1152340	38,2	10,2
Bikini top	841	1126193	33,2	11,5
Swimwear bottom	1275	1104215	35,4	10,7
Underwear bottom	2600	1068059	37,1	10,8
Skirt	2591	934252	36,2	11,4
Shirt	3232	787733	38,6	12,3
Leggings/Tights	1749	734119	38,0	9,9
Jacket	3646	599933	38,8	11,2
Socks	1773	483356	37,2	10,6
Hoodie	2205	483282	38,8	11,2
Blazer	1072	448100	38,3	12,3
Cardigan	1430	380077	39,8	11,7
Swimsuit	637	285910	35,4	9,8
Jumpsuit/Playsuit	1094	227629	35,7	10,0
Bag	1214	218837	35,9	12,2
Belt	435	202416	35,9	12,1
Underwear Tights	510	201901	37,7	9,4
Pyjama set	1071	192233	39,3	10,0
Boots	943	182865	37,6	11,1
Earring	1112	175526	34,0	11,6
Scarf	929	156463	39,5	12,2
Bodysuit	873	138315	34,5	9,3
Necklace	552	131665	33,9	12,0
Sandals	728	125262	36,8	10,6
Sneakers	1542	124267	38,9	10,2
Coat	429	120481	38,8	12,7
Sunglasses	600	117553	34,5	11,2
Unknown	113	97040	36,9	12,2
Pyjama bottom	213	95732	39,2	12,4
Other accessories	960	92816	36,0	11,5
Hair/alice band	808	90656	35,6	11,7
Polo shirt	428	88600	40,6	12,4
Hat/beanie	1201	87690	37,2	10,4
Underwear body	169	85338	33,6	10,7
Heeled sandals	197	62680	37,1	12,2
Ballerinas	355	49261	38,5	11,0
Night gown	159	47436	39,6	11,1
Flat shoe	156	47144	36,8	12,3

Hair ties	24	45575	34,0	11,9
Hair string	226	44094	35,5	11,4
Sarong	65	43389	38,0	12,4
Hair clip	229	40053	36,3	11,8
Garment Set	1219	39614	35,8	9,9
Wedge	110	38657	38,6	11,7
Ring	237	37364	32,6	11,5
Dungarees	292	35101	36,3	9,9
Gloves	332	35028	38,9	9,9
Robe	126	34948	37,3	10,9
Pumps	183	32580	39,4	11,9
Cap/peaked	528	31068	36,9	10,1
Other shoe	369	26991	37,9	10,9
Hat/brim	376	26941	37,3	10,5
Slippers	240	23550	37,5	10,5
Flip flop	119	21299	37,8	10,2
Swimwear set	184	19119	40,4	8,5
Pyjama jumpsuit/playsuit	361	12700	36,8	9,4
Underwear set	46	12530	36,9	11,2
Outdoor Waistcoat	140	10361	40,0	10,8
Bracelet	164	10252	36,8	12,1
Earrings	11	10006	31,8	12,0
Costumes	83	9658	39,4	9,2
Kids Underwear top	90	9214	43,9	7,5
Underdress	20	8694	39,1	12,8
Outdoor trousers	115	7938	40,4	8,8
Heels	22	6952	32,3	11,1
Watch	71	6530	34,8	12,3
Tie	131	6452	37,4	11,6
Nipple covers	19	6261	33,8	11,6
Tailored Waistcoat	71	5752	39,5	10,6
Umbrella	26	5047	36,7	12,9
Beanie	50	4919	34,2	10,5
Wallet	75	4754	34,9	12,2
Dog Wear	20	3554	38,3	13,4
Hairband	2	2801	35,5	13,8
Tote bag	2	2754	31,9	12,8
Underwear corset	7	2703	33,6	9,2
Bootie	28	2639	34,6	8,9
Waterbottle	21	2172	34,4	10,6
Braces	3	2088	35,4	13,4
Bucket hat	7	2066	32,4	12,5
Outdoor overall	53	1884	36,3	9,5
Mobile case	4	1835	31,3	12,0
Backpack	6	1731	33,2	12,7
Bra extender	1	1536	38,5	14,0
Cap	13	1526	35,2	11,6
Felt hat	9	1486	32,9	9,9
Straw hat	6	1424	34,4	11,0
Weekend/Gym bag	9	1398	36,4	11,9
Fine cosmetics	41	1396	39,7	10,5
Soft Toys	41	1383	41,5	11,0
Long John	25	1194	41,1	8,7
Flat shoes	10	1111	37,5	13,7
Alice band	6	1061	34,9	12,1
Cross-body bag	5	909	36,0	13,7

Side table	13	533	37,3	12,2
Shoulder bag	2	505	35,0	13,5
Sleeping sack	42	473	36,4	9,8
Leg warmers	6	321	40,4	10,2
Swimwear top	45	314	38,9	4,8
Giftbox	14	282	38,1	10,7
Dog wear	5	265	37,6	12,1
Wireless earphone case	2	265	29,5	11,8
Marker pen	5	229	32,3	12,3
Baby Bib	3	170	33,8	8,2
Headband	1	154	41,6	15,0
Keychain	1	123	35,1	13,4
Sewing kit	1	105	29,7	9,8
Chem. cosmetics	3	100	37,4	11,0
Sleep Bag	5	80	37,1	7,5
Zipper head	3	64	43,5	13,2
Accessories set	5	48	31,9	7,4
Moccasins	4	41	37,4	11,5
Stain remover spray	2	36	38,9	13,4
Washing bag	1	35	37,3	13,0
Wood balls	1	35	43,5	10,2
Blanket	1	28	34,7	7,2
Cushion	1	26	37,6	7,4
Eyeglasses	2	25	31,9	13,1
Towel	1	20	32,9	8,4
Bumbag	1	16	34,0	14,8
Clothing mist	1	4	45,8	5,7
Toy	1	4	32,0	9,8

Table 9 Summary Statistics per Color

The table provides summary statistics for product color variable used in this study. The table reports the number of observations (N), total number of customers, the mean of average age and the mean of standard deviation of age for all 52 unique product colors in the dataset.

Colors	N	Customers	Mean of Avg. Age	Mean of St. Dev. Age
Black	21566	11036124	3,7	11,4
White	9149	3367903	3,7	11,3
Off White	2618	841582	3,7	11,6
Light Beige	3223	1263294	3,7	11,8
Beige	2610	921122	3,7	12,0
Grey	4194	839813	3,9	10,5
Light Blue	2900	913390	3,7	11,1
Light Grey	1959	404972	3,8	10,9
Dark Blue	11505	2179995	3,9	10,5
Dark Grey	2550	716897	3,9	10,9
Pink	1921	492063	3,6	10,8
Dark Red	2207	717285	3,7	11,3
Greyish Beige	214	60187	3,9	11,5
Light Orange	1452	384033	3,6	11,0

Silver	660	117300	3,4	11,8
Gold	1325	244001	3,4	11,5
Dark Pink	782	191714	3,7	11,1
Yellowish Brown	1392	413974	3,8	11,9
Blue	3186	1088347	3,8	10,8
Light Pink	5440	857905	3,7	10,8
Light Turquoise	961	72440	3,8	9,7
Yellow	1565	429232	3,7	11,0
Greenish Khaki	2612	761277	3,9	11,1
Dark Yellow	543	114472	3,7	10,5
Other Pink	693	71694	3,9	9,7
Dark Purple	285	50364	3,8	11,6
Red	2850	768462	3,7	11,2
Transparent	27	8439	3,4	11,0
Dark Green	1982	713879	3,8	11,2
Other Red	106	15362	3,7	10,6
Turquoise	411	72259	3,8	10,3
Dark Orange	846	276759	3,7	11,0
Other	92	21014	3,7	10,8
Orange	739	163773	3,7	11,2
Dark Beige	1025	290662	3,8	11,9
Light Green	635	122963	3,7	10,8
Other Orange	144	27788	3,8	10,3
Purple	172	28889	3,6	11,2
Light Red	269	33465	3,7	10,1
Light Yellow	927	185644	3,7	11,0
Green	758	187722	3,7	10,4
Light Purple	519	108733	3,5	11,3
Dark Turquoise	436	117819	3,8	11,1
Other Purple	43	7868	3,5	11,0
Bronze/Copper	88	9864	3,9	10,2
Other Yellow	217	41901	3,7	10,2
Other Turquoise	12	553	3,8	9,1
Other Green	125	14890	3,9	9,7
Other Blue	47	8053	3,8	10,4
Unknown	28	5595	3,7	12,8

Table 10 Summary Statistics for Perceived Colors

The table provides summary statistics for product perceived color variable used in this study. The table reports the number of observations (N), total number of customers, the mean of average age and the mean of standard deviation of age for all 8 unique perceived product colors in the dataset. These colors are defined by H&M based on how the customers perceive the actual color of the product. The detail explanation of this variable is available in Section 3.

Color	N	Customers	Mean of Avg. Age	Mean of St. Dev. Of Age
Dark	40427	15587748	3,8	1,1
Dusty Light	20954	5404352	3,7	1,1
Light	14989	4573170	3,7	1,1
Medium Dusty	12048	3755113	3,8	1,1
Medium	5393	1231702	3,8	1,1
Bright	6079	1205042	3,7	1,1
Undefined	92	21014	3,7	1,1
Unknown	28	5595	3,7	1,3