Supply chain planning and inventory management | Retail | Predictive forecasting | Clustering

Data-driven clustering for new garment forecasting

INDITEX

BUSINESS PROBLEM

The ability to detect patterns early in the design process is critical for fashion firms to make decisions, particularly given the speed at which new garments are introduced. Traditionally, most garment defining features were only used by designers and buyers in and since the data was intractable for a computer: shape, color, fit, etc. In particular, deciding on the size-curve distribution (percentage of smalls, mediums and larges) for a new garment relies heavily on finding a comparable, i.e. a previous garment that is similar to the previous one. Could this process be automated by using non-traditional data like the one described above?

APPROACH

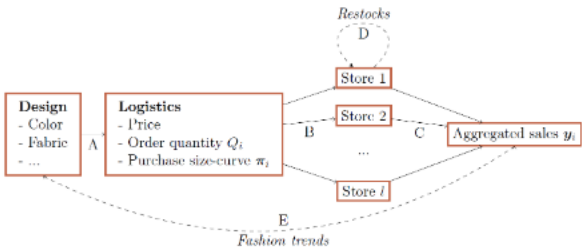
By using NLP techniques that preserve semantics we can extract the garment characteristics and create numerical embeddings of the garments. In tandem, we develop two custom algorithms for the fashion industry that leverage the dataset above to automate the comparable finding and simultaneously forecast the size-curve distribution.

DATA SOURCES

All the data comes from structured internal data sources that are easily queryable. However, there was a lot of data processing and wrangling to ensure that it could be used by the algorithms. In particular, the images and garment characteristics that were already encoded in a proprietary text format.

Data Types and Format

Rectangular data for time series of sales and purchases broken down by garment and size. Garment images for visual comparisons. Text descriptors that represent unique garment characteristics.



Author: Gianpaolo Luciano Rivera

IMPACT

The dataset itself: i.e. a vectorial embedding of the garments that preserves the defining is intrinsically valuable for a fashion firm. These embeddings are useful for computationally defining what similarity means in a sense that replicates what humans understand. Therefore, it has several applications, from item recommendations, cold-start clustering and performance indicators and forecasting tasks for new garments. More importantly, by using this data and framework we are automating the comparable process finding which can save hundreds of hours to the designers & buyers. Last, we show that the two forecasting algorithms can achieve close to human level results for the size-curve problem. Given enough data (to leverage as known comparables) it should improve baseline human performance.

DRIVERS	Understanding the established procedures and designing a solution that worked with the designers and buyers. Seeking to leverage their experience and know-how of the business in finding comparable garments. We weren't looking to replace them but to enhance their job and make it more seamless. In addition, the macro-trends of the industry that require even more expedited end-to-end (design-purchase-selling) experiences for the final consumers.
BARRIERS	Data quantity: the dataset test is just a small subset of what's available. To achieve human level results on the forecast, the algorithms need a complete set of all the garments of the firm. Data quality: the initial text representations of the garments might not be as granular to clearly capture all the details of the garments themselves which are critical for human level comprehension.
ENABLERS	The fact that Zara is a pioneer in the usage of data in the fashion industry, particularly in their supply chain & distribution processes. The data was readily available, albeit incomplete due to technical reasons. Similarly, the technical prowess of their analytics and data science teams is unrivaled.
ACTIONS	1) Interviews with the clients (buyers and designers of the garments) 2) Data data mining/processing to ensure the consistency across several seasons 3) Creation of a python package that queries/processes and standardizes the data. This package is extensible enough so that several machine learning models can be used in the forecasting algorithm easily
INNOVATION	1) Using a custom Natural Language Processing pipeline to embed the unique text characteristics of the garments that preserve hierarchical clusters, semantics and human level similarity 2) Development of two algorithms: Cluster-While-Regress (CWR) and k-nearest neighbors (kNN) algorithms tailored for the size-curve problem using the data above
IMPROVEMENT	Due to the data problem described above, there wasn't a quantifiable improvement in the operations of the firm at the moment. In the future when the framework and algorithms are improved, we will be able to give a concrete result that will come in the shape of savings.
BEST PRACTICES	Making sure that there's enough data quantity: i.e. sufficient number of garments. And, data quality: ensuring that the text descriptors are sufficiently descriptive to capture all the details of the garments.
OTHER APPLICATIONS	The data structures can be used for: item recommendations, cold-start clustering, performance indicators and forecasting tasks for new garments. The algorithms can be used for any forecasting task