Leveraging the Power of Images in Managing Product Return Rates

Daria Dzyabura*

New Economic School, Moscow, Russia, ddzyabura@nes.ru

Siham El Kihal

Frankfurt School of Finance & Management, Germany, s.elkihal@fs.de

John R. Hauser

MIT Sloan School of Management, USA, hauser@mit.edu

Marat Ibragimov

MIT Sloan School of Management, USA, mibragim@mit.edu

Marketing Science 2023 Vol. 42, No. 6, 1125-1142.

Prepublication version. Published version available from INFORMS.

*Authors listed alphabetically

This paper has benefited from feedback obtained from Marketing Effectiveness through Customer Journeys and Multichannel Management, Bologna, June 2019, the JAMS Thought Leaders' Conference on Innovating in the Digital Economy, Bocconi University, June 2019, the 11th Triennial Invitational Choice Symposium, Cambridge, MD, June 2019, the 48th Conference of the European Marketing Academy (EMAC), University of Hamburg, June 2019, the Theory + Practice Conference, Columbia University, NY, May 2019, the Annual Marketing Research Camp, Tuck School of Business, Dartmouth, May 2019, the 47th Conference of the European Marketing Academy (EMAC), University of Strathclyde, Glasgow UK, the Special Session on Machine Learning, 40th INFORMS Marketing Science Conference, Fox School of Business, Temple University, USA, the Workshop on Multi-Armed Bandits and Learning Algorithms, Rotterdam School of Management at Erasmus University, Rotterdam, May 2018, the Washington University Foster School of Business Seminar Series, February 2018, and the NYU 2017 Conference on Digital, Mobile Marketing, and Social Media Analytics, NYU Stern School of Business, New York. We also wish to thank Martin Artz, Katrijn Gielens, and Katharina Hombach for comments and suggestions.

Leveraging the Power of Images in Managing Product Return Rates

In online channels products are returned at high rates. Shipping, processing, and refurbishing are so costly that a retailer's profit is extremely sensitive to return rates. Using a large dataset from a European apparel retailer, we observe that return rates for fashion items bought online range from 13% to 96%, with an average of 53% – many items are not profitable. Because fashion seasons are over before sufficient data on return rates are observed, retailers need to anticipate each item's return rate prior to launch. We use product images and traditional measures available prelaunch to predict individual item return rates and decide whether to include the item in the retailer's assortment. We complement machine-based prediction with automatically extracted image-based interpretable features. Insights suggest how to select and design fashion items that are less likely to be returned. Our illustrative machine-learning models predict well and provide face-valid interpretations – the focal retailer can improve profit by 8.3% and identify items with features less likely to be returned. We demonstrate that other machine-learning models do almost as well, reinforcing the value of using prelaunch images to manage returns.

Keywords: machine learning, image processing, deep learning, product returns

1. Introduction

Online retailers are challenged by the high cost of product returns. Processing and refurbishing the returned item is so costly that large retailers such as Amazon and Walmart allow customers to keep the item, because it often costs more to ship and process the returned product than the product is worth (Wall Street Journal 2022). Nick Robertson, founder of the UK's largest fashion retailer, ASOS, stated that a 1% drop in ASOS' return rate could increase the firm's bottom line by an impressive 30% (Thomasson 2013).

In the $500 billion fashion industry, return rates are high, and vary greatly by item. The products upon which we focus are fashion items. For our focal retailer, a large European apparel retailer, we observe item return rates averaging 53% ranging from 13% to as high as 96% for some items. This is in contrast with the 3% return rate in the same retailer's offline channel, with the same set of items. Even with high margins, the items on the higher end of this return-rate spectrum generate a net loss for the firm's online store. In fashion, as in many industries, the product return rate is key input into any product management strategy. The problem in the fashion industry is that fashion seasons are short and return deadlines are generous. By the time an item's return rate is observed, the fashion season is well underway or almost over. To effectively manage item assortment in light of returns, it is critical that the retailer is able to predict item return rates using only data available prelaunch.

In this paper, we address this problem by leveraging image processing methods. We demonstrate that item images improve predictions of return rates, that policies based on predictions can improve profit, and that data-based insights are face valid, internally consistent, and suggest which items are returned at high and low rates. To do so, we develop a modeling framework to predict and interpret how product images relate to their return rates. Machine learning models produce accurate predictions of an item's return rate based on features of the product image and other characteristics available prelaunch. For example, including deep-learning image features in gradient-boosted regression trees

(GBRT) predicts 13.5% better than a model based on traditional features alone. Using this model and the derived policy to decide on which items to display results in a profit improvement net of returns by 8.3% relative to displaying all items in the online channel. SHAP values (that relate automatically-interpretable image-based features to return rates) suggest how the firm might design (or otherwise source) items less likely to be returned.

We tested a variety of alternative machine-learning models and features to suggest which do well and which do not on our data. Among those tested are deep-learning features, human-coded features, hand-crafted automated pattern and color features, and automatically-generated image-based interpretable features. We find that many machine-learning models do well on our data providing evidence for the value of item images for managing item assortments.

Our contribution is to show that incorporating item images into models helps a firm decide, prior to launch, which products to include in its online store based on profitability net of returns. The approach is fully automated, scalable, and implementable prior to product launch, and an improvement on current practice that does not incorporate product images. The approach has the advantage that it can be easily implemented by a retailer for each fashion collection.

The remainder of the paper is organized as follows. We begin by reviewing relevant literature on managing product returns and leveraging image data (§2). §3 describes the data and provides empirical (model-free) evidence that image data predict returns. §4 demonstrates that image-based features improve predictions, explores alternative models, and develops a model-based policy for selecting which items to display/not-display in the online store. §5 complements the predictive model with automatically-generated image-based interpretable features which provide insights on how to source and design items. We conclude with a summary, limitations, and suggested future research (§6).

2. Related Literature

We build on and contribute to two streams of literature: managing product returns and

leveraging image data.

2.1. Managing Product Returns

A rich literature in marketing and operations investigates firm strategies for managing product returns. One such strategy is to manage returns by optimizing the leniency of the return policy (such as fees, prices, or deadlines). Anderson et al. (2009) develop an individual-level model of purchase and return and use it to optimize the return costs for customers. Moorthy and Srinivasan (1995) suggest that a return policy is a signal of item quality. Shulman et al. (2011) show that the optimal policy (strict vs. lenient) balances sales and returns. See also Davis et al. (1998), Wood (2001), Bower and Maxham-III (2012), and Janakiraman et al. (2016).

Another approach for managing returns focuses on managing customers and understanding their return behavior. For example, Petersen and Kumar (2009, 2015) describe customer return behavior and how it affects future spending and how this can be accounted for in lifetime value calculations to target more profitable customers. Sahoo, Dellarocas, and Srinivasan (2018) study how product reviews decrease return rates by reducing consumers' uncertainty. Other studies link product returns to factors such as prices and price discounts, marketing instruments, free shipping promotions, the use of an app, or even the weather (e.g., Conlin et al. 2007, El Kihal et al. 2021, Narang and Shankar 2019, Petersen and Kumar 2009, 2010, Shehu et al. 2020, El Kihal and Shehu 2022). Other than at the broad category level (Hong and Pavlou 2014), the literature has not explored the characteristics of products related to high return rates.

Rather than focusing on return policies, managing (and "firing") customers, or prices and marketing strategies, our research focuses on the products (items) themselves. Not only is this a gap in the literature, but it is clearly complementary to return policies, managing customers, prices, and marketing strategies. We focus on which items to display/not-display and item features that lead to high or low return rates. Not displaying an item is mathematically equivalent to charging an infinite price.

With new data, future research might explore softer strategies such as setting a very high price. Pricing research is not feasible with our data and beyond the scope of this paper. We observe prices and use prices as traditional product features, but we do not observe the demand curve and cannot optimize prices.

2.2 Leveraging Image Data

Images have always been an important part of firms' marketing efforts. Over the past two decades, technical advances and the rise of digital platforms have created an abundance of visual data. Together with the development of image processing tools and advanced modeling techniques, these data have created unique research opportunities in marketing. For example, researchers have analyzed images in consumer reviews (Zhang and Luo 2022), user-generated digital content (Hartmann et al. 2021; Liu et al. 2018; Klostermann et al. 2018; Dzyabura and Peres 2021), firm logos (Dew et al 2022), and seller images on digital platforms (Zhang et al 2021). For a detailed review of published and ongoing research, see Dzyabura et al. (2021).

He and McAuley (2016), Lynch et al. (2015), and McAuley et al. (2015) demonstrated by example that images are valuable for making recommendations regarding clothing styles, substitutes, and personalized rankings. Shi et al. (2021) use machine learning to identify garments and classify fashion-item features from street snapshots, runway photos, and online stores. They use these tools to interpret fashion dynamics and conclude that machine learning can identify fashion features not discussed in fashion magazines. They do not use the fashion features to predict item sales or item returns.

The literature supports that images contain valuable information in many product categories, and specifically in fashion. Although images have not been used to forecast return rates for specific items, they have been proven valuable for other tasks. The literature also suggests that machine learning can identify independent variables ("features") that are used to forecast dependent variables (in our case return rates for each item).

3. Data Description and Empirical Evidence that Image-based Data Predict Return Rates

To study product returns, we chose an industry that is particularly challenged by high return rates – women's apparel. We obtain transaction data from a major European mono-brand retailer-manufacturer. We augment the transaction data with a study in which human judges label images from the retailer's two largest categories (dresses and shirts).

3.1. Retailer Transaction Data

The women's apparel retailer has a network of 39 retail stores in Germany complemented by a large online operation that accounts for 30.5% of its sales. All items appear in both channels and are always sold for the same price in the two channels. The retailer has a lenient return policy mandated by law: customers can return any purchased item for any reason within 14 days, without providing the reason. By the retailer's policy, items must be returned in the same channel in which they were purchased.

We use data on 1,231,055 transactions, including sales and returns, that occurred in online channels during the observation period from September 1st, 2014 until August 31st, 2016 (two full consecutive years). We exclude non-apparel items such as perfume, gift cards, or accessories. We observe returns for all orders made within the observation period. The data also include offline sales but returns are rare in the offline channel (with a 3% offline return rate for the focal retailer).

For each transaction, we observe the date, the channel (online/offline), item identifiers, and which items were returned. For each item, we observe the category (e.g., dresses), price, and four-to-six images. The images were taken by the same studio, using standardized procedures, resulting in consistent image quality. In our primary analysis, we include only the front image of each item, which is the most informative and always the first image displayed to the customer on the retailer's website. (A model using all images performs only marginally better – Online Appendix B.) The images display the item by itself (not on a model or a manikin) against a white background. We include only items for which

we have images (97% of items) and which were sold at least 20 times. (Varying this threshold did not change our findings – Online Appendix B.) On average, an Item fills 55% of the image (standard deviation 13%). The fill-rate mostly depends on the item category (for example, dresses are likely to be larger than shirts). Among all the item images, 99% have the same size (2200 x 1530 pixels). Figure 1 contains example images of four items.

Figure 1. Examples of Images of Four items

(a) (b) (c) (d)

The resulting data contain 4,585 distinct items from fifteen different apparel categories, as categorized by the retailer. Return rates for items sold via the online channel lie within 13–96% (56% average for this subsample, slightly above the overall average of 53%). These rates are well above return rates in the offline channel (3%) where consumers can touch, feel, and try on items. The processing and refurbishing costs of returns from the online channel, our focus, is well above returns from the offline channel (return-cost data are proprietary). For the interested reader, data on consumer characteristics complement our focus on displaying/not-displaying and designing items. For example, evening shoppers return more than morning and daytime shoppers; middle of the week shoppers return less than beginning and end-of-the-week shoppers; and price discounts are positively related to return rates. These data are only observed postlaunch and cannot be used to manage item returns prelaunch, hence postlaunch data are not used by our models. For completeness we provide these data in Online Appendix B.

3.2. Data Augmentation with Human-Coded Features (HCF)

The number of items per fashion season is large and fashion seasons change rapidly. We seek an automated way for the firm to manage assortments. It would be prohibitively expensive, and the firm may not have the time between fashion seasons, to ask humans to code the item images. Nonetheless, to explore whether or not image data are sufficient without human-coded data, we asked human judges to code illustrative fashion features in the two largest categories. These data provide evidence that image features are predictive of return rates and provide a benchmark with which to evaluate the predictive ability of the automatically-generated image-based interpretable features studied in §5. The automatically-generated fashion features are curated to apply across categories and, hopefully, subsume many specific fashion features such as the human-coded features for dresses and shirts.

We conducted a study in which human judges labeled 2,392 images from the largest two categories of items (dresses and shirts). Four independent judges, blind to the purposes of the study, labeled each clothing item with respect to symmetry (symmetric vs. asymmetric), pattern (solid, floral, striped, geometric/abstract), and additional details (text, metallic/sequin, graphic, lace). Three of the judges coded sleeve length (short, medium, long, sleeveless), and the presence of belts and/or zippers. The human-based label for an image is equal to 1 if the majority of the judges indicate attribute presence, and equal to 0 otherwise. Ties were uncommon and broken by across-item percentages. On average, judges agreed with the majority vote for 91.7% of the judgments.

3.3. Model-Free Motivation: Observable Variables and Image Data Relate to Return Rates

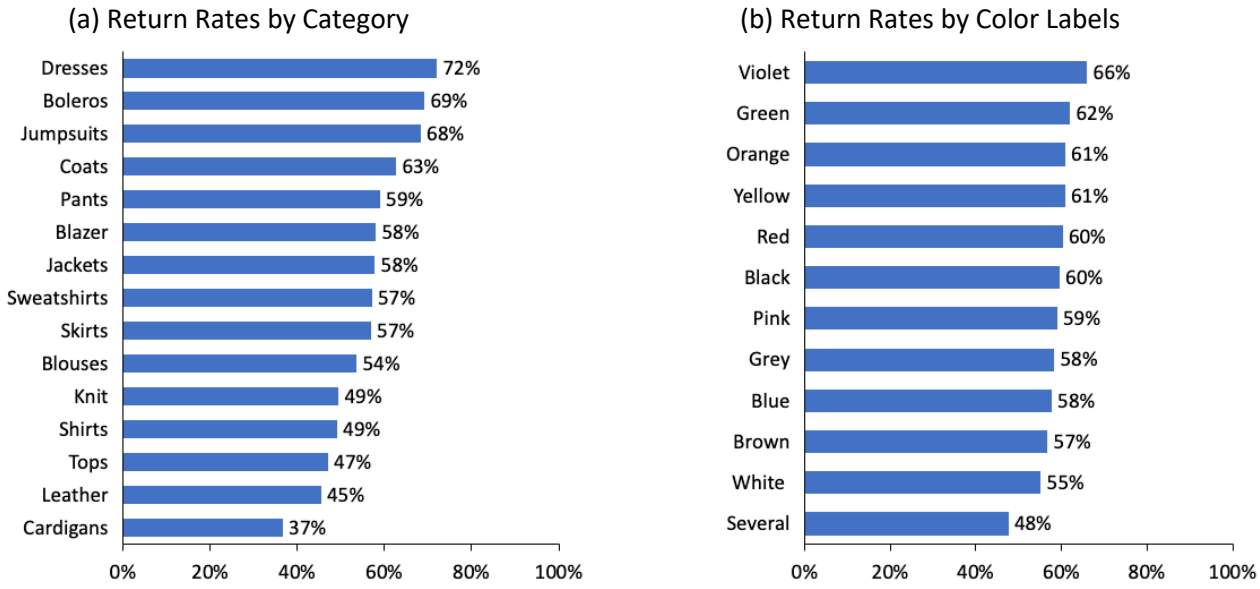
Non-image variables. Return rates are related to observable variables. For example, return rates vary by category as illustrated by Figure 2a. Dresses, the largest category in our data, are returned on average 72% of the time while cardigans are returned 37% of the time. Seasonality and price are the other variables used by the retailer that are related to returns (evidence in Online Appendix B).

Color. A minimal use of images is the color of the item. For example, consumers can more easily

imagine themselves in common conservative colors such as blacks, blues, and greys, but often want to try fashion colors such as pinks, purples, and pastel colors. To examine whether return rates vary by image data, we begin with the color labels (twelve color bins) that the retailer uses to categorize each item. The bins are not perfect, for example, "pink" includes many shades of pink and a single color does not fully summarize multi-colored patterns or highlights. Nonetheless, Figure 2b suggests that color labels are related to return rates.

We will show that color labels augment traditional models based on category, price, and seasonality. We will also demonstrate that we can do even better with more comprehensive image features and models that account for interactions and non-linearities (and that are regularized).

Figure 2. Online Return Rates by Category and Color – Retailer's Traditional Classifications



Human-coded features (HCFs). HCFs are not designed to be scalable to all items in all categories for every fashion season. But they are valuable as indicators that image features are related to return rates. Figure 3 displays the correlations of the HCFs with return rates. Asymmetrical items are associated with higher return rates, compared with symmetrical items. Items with patterns (floral, striped, or geometric/abstract) have lower return rates compared with items without a pattern (solid items).

Among the additional details, lace details, metallic/sequin details and belts seem to be associated with higher return rates, while the presence of a zipper and text or graphic details is associated with lower return rates. Finally, the length of sleeves is negatively correlated with the return rate. When we regress the item return rate on the HCFs features, we get similar insights. Details are in Online Appendix B.

Figure 3. Relationship between the Human-coded Features (HCFs) and Return Rates

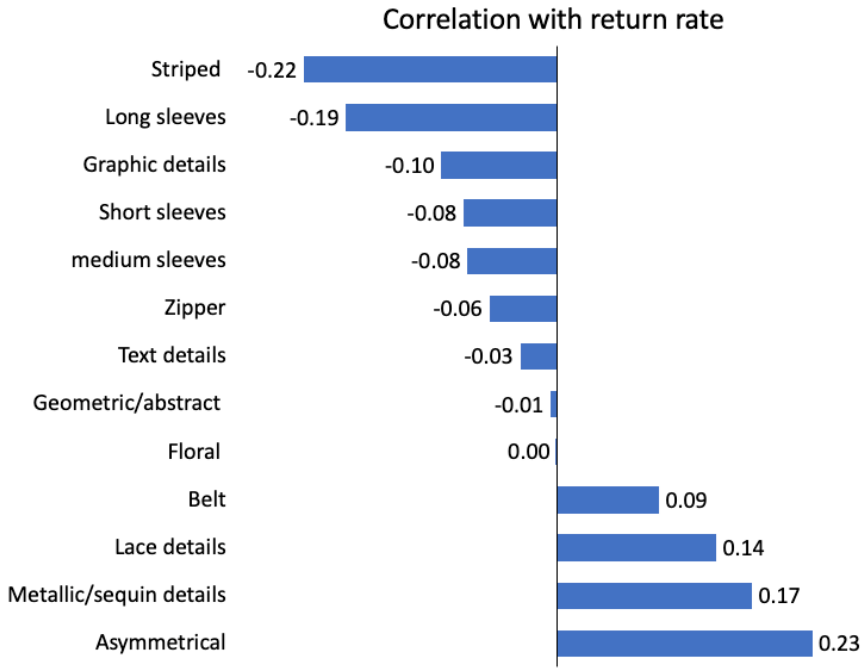


Figure 3 motivates the hypothesis that image-based features relate to return rates. Likely the relationship is more complex than simple correlations—the HCFs likely interact with each other and with the traditional measures such as fashion category. In the next section, we explore models to handle complex interactions. Because the HCFs do not scale, §5 develops more general and more comprehensive automatically-generated image-based interpretable features. Model-free evidence, §5.2, confirms that the interpretable features are correlated with return rates. §5 also explores model-based methods, called SHAP values, that relate the interpretable features to return rates.

4. Prediction: Using Prelaunch Image Data to Predict Return Rates

The previous section suggests that images (image features) augment traditional measures when

predicting return rates. We seek a good predictive model to support the retailer's decisions on which items to sell in its online store (display/not-display). For the retailer's policy, we focus on the profitability of individual items rather than the number of units per se. If an item is not profitable, then the retailer's decision whether or not to sell it does not depend upon the forecasted number of units sold. Consistently with the managerial goal, we summarize return behavior by a return rate for each item. The return rate per item varies between 0% and 100%.

Let r_i be item i's return rate, defined as the ratio of the number of returned units ($N_{i,returned}$) to the number of purchased units of the item ($N_{i,purchased}$):

$$r_i \equiv \frac{N_{i,returned}}{N_{i,purchased}}. \quad (1)$$

To manage product returns using predictions of r_i we make three modeling decisions. The first decision is which features to extract from the images. The second decision is which model to use to predict r_i as a function of the image features and the traditional variables. The third decision is the display/not-display policy used by the retailer—the policy is a function of the return rate and the model's predictive ability. We begin by defining the criteria we use to evaluate predictive models.

4.1. Criteria to Evaluate Predictive ability

Our primary criterion is the *out-of-sample* R^2_{model}, calculated on all items (i) in our sample (K_{all}). For ease of presentation, we multiply R^2_{model} by 100.

$$R^2_{model} = 1 - \frac{\sum_{i \in K_{all}} (r_i - \hat{r}_i^{model})^2}{\sum_{i \in K_{all}} (r_i - \hat{r}_i^{average})^2}, \quad (2)$$

where $\hat{r}_i^{model}$ is the out-of-sample item return rate predicted by our model. We use twenty-fold cross-validation to generate out-of-sample predictions for each point in a sample. We randomly divided our sample into twenty non-overlapping folds, where we used 75% of folds to train the model, 20% to validate the model (optimized over a set of hyperparameters, Appendix), and the remaining 5% to compute out-of-sample predictions. By assigning different folds to training, validating, and testing the

model, the cross-validation procedure allows us to construct out-of-sample predictions for all points in the sample.

To ensure reliable estimates of item return rates, we exclude items that were sold fewer than twenty times. We obtain the same results if we screen the data to require a minimum threshold of either 10 or 30 unit sales. The retailer's decisions are item-by-item because in the online store items are displayed singly and interactions are uncommon. This is one way in which managerial decisions differ from the offline store where items are displayed together and interactions are common. The models do not improve when we use $N_{i,purchased}$ to weight the data for each item. Details in Online Appendix B.

Because other policies might depend on other criteria, we examine the robustness of our methods to different performance measures: we supplement R^2_{model} with mean absolute deviation (MAD) and U^2_{model}. U^2_{model} is a common measure used in marketing that is based on information theory (and probabilities) and measures the amount of (Shannon's) information explained by the model relative to that explainable by perfect predictions (Hauser 1978). Although derived for classification (0 vs. 1), U^2_{model} applies to more continuous measures such as r_i. It differs from R^2_{model} and MAD because it uses logarithms rather than squared or absolute error. Although derived from information theory, U^2_{model} is sometimes called a pseudo-R^2. Other classification metrics, such as area under the curve (AUC), are derived for a 0 vs. 1 outcome. The extension of AUC to continuous measures is proportional to MSE and would be redundant with R^2_{model} (Hernández-Orallo 2013, Theorem 7 & Corollary 8).

4.2. Baseline Predictions (Item's Category, Seasonality, and Price)

Before we explore the use of images to manage returns, we explore non-image baseline predictions that use information routinely collected by the retailer. For each item in its inventory, the retailer observes the seasonality (month), the item category (e.g., dresses), and price. For price, we use the average price at which the item was sold. Other measures, such as price relative to average category price, do not improve predictions. For the purposes of this analysis, we treat price as exogenous to the

decision on whether or not to display an item in the online store. Our data do not contain sufficient information on the demand curve to optimize price. We demonstrate that profitability is improved when price is exogenous. Future research with improved data could include price optimization in policies to improve profitability further.

We can choose a variety of prediction models with which to predict return rates as a function of image and non-image features. These methods vary from simple regression to highly nonlinear functions obtained with machine learning. In our data, we obtain the best predictive ability using gradient boosted regression trees (GBRTs). Bagging methods (random forest) and LASSO do not predict as well as a GBRT, although image-feature-based models using these methods provide incremental predictive ability relative to models based on non-image features alone. Details are in Online Appendix B.

Table 1 reports the predictive ability of the baseline model. To address the variance in the estimated R^2_{model} due to randomness of the division into folds, we generated twenty-five different sets of cross-validation folds (each set including twenty folds); we report the average and standard deviations of the estimated R^2_{model}.

4.3. Improving Predictions with Images (Baseline Plus Color Labels)

Empirical model-free evidence in §3.3 suggests that color labels are related to return rates. Color labels are minimal image-based features and can be used without image-processing. Table 1 shows that color labels improve predictions slightly relative to the baseline. While the improvement is small, the color-label model is further evidence that there is information in images. We show next that deep-learning image features improve predictions substantially beyond predictions obtainable with the retailer's color labels.

Table 1. Baseline and Color-Label-Model Predictions

Model	Non-Image Features	Image Features	Out of Sample $U^2 * 100$	Out of Sample MAD $* 100$	Out of Sample R^2	R^2 Improvement over baseline
Non-Image Baseline	Category, seasonality, price	None	52.75 (0.27)	8.59 (0.01)	41.31 (0.18)	–
Color-labels added to baseline	Category, seasonality, price, and color labels	None	53.56 (0.28)	8.48 (0.01)	42.50 (0.20)	2.88%

Note: Models use LightGBM and differ only with respect to the set of features included. Standard deviations are reported in parentheses.

4.4. Predictions Using Deep-Learning Image Features

Images are more than just color. Consider the three items in Table 2. The first item, the white top, is easily categorized and a common color; the color-label model does well. The second item, the top with stripes, is multicolored and hard to categorize by color; the color-label model does less well. The third item, the dress, is readily categorized as pink, but the color-label model does not do well, likely because the pink is not a prototypical pink and because the dress's shape does not work well for everyone.

Table 2. Return Rates and Color-Label Predictions for Three Apparel Items

			
Actual Return Rate minus Color-label Prediction	+1.0%	– 12%	+15.2%

Note: Actual return rates are not included for confidentiality reasons.

To improve upon the color-label-based benchmark, we examine image-processing features

identified with a convolutional neural network (CNN). In our data, CNN-based features predict better than other image-based features. We examine the predictive ability of other image-based features in §4.5 and more interpretable automatically-generated image-based features in §5.

The information in apparel images is more complex as illustrated by the shape of the pink dress in Table 2. Other dresses might feature floral patterns or complicated geometric shapes. Deep-learning algorithms have the advantage that they learn feature representations automatically and can be modified for particular applications. To explore the potential of deep learning for image-based predictions of apparel return rates, we use an established CNN. Through a series of nonlinear filters and transformations, the CNN learns highly complex nonlinear transformations to map an image to a set of deep-learning features. The tradeoff is that, while good for prediction, the CNN features are difficult to interpret. The CNN features likely capture the information provided by more-specific features (including the HCFs), but without interpretability, we do not gain insight into which features are associated with high return rates. For greater detail on each transformation and for an application of a CNN to unstructured marketing data, see Zhang and Luo (2022).

Our 4,585 images are not sufficient to train a deep CNN from scratch, thus we use the second-to-last pre-output layer of the Residual Neural Network (ResNet; He et al. 2015). ResNet won the 2015 ImageNet Large Scale Visual Recognition Challenge and was trained on the ImageNet data set (1.3 million images in roughly 1,000 categories). The ResNet network has 152 layers, making it one of the deepest networks yet presented on ImageNet. The second-to-last layer of the network contains 2,048 features. (The last layer is the output layer.) The 2,048 deep-learning features were used directly in the GBRT.

In Table 3, we see substantial improvement when using deep-learning features relative to the baseline and color-label models. This improvement in predictive ability leads to a substantial improvement in profit when using the display/not-display policy (§4.6). Because the 2,048 deep-learned features are likely to encode well the image information, we expect little or no improvement when we

add other machine-learned features to the deep-learning features (see next section). Returning to the images in Table 2, the GBRT based on deep-learning features predicts return rates better for the hard-to-predict items. The GBRT/CNN model predicts a return rate for the striped top within 4.7% of the true rate (color-label predictions are within 12%), and a return rate within 5.6% for the pink dress (color-label predictions are within 15.2%).

Table 3. Predictions Using Deep Learning Image-Processing Features

Model	Non-Image Features	Image Features	Out of Sample $U^2 * 100$	Out of Sample MAD $* 100$	Out of Sample R^2	R^2 Improvement over baseline
Non-Image Baseline	Category, seasonality, price	None	52.75 (0.27)	8.59 (0.01)	41.31 (0.18)	–
CNN Features	Category, seasonality, price, color labels	Deep-learning	56.70 (0.26)	8.15 (0.02)	46.88 (0.19)	+13.48%

Note: Models use LightGBM and differ only with respect to the set of features included. Standard deviations are reported in parentheses.

The ResNet CNN is not the only image-processing model that does well on our data, but it is the best of those tested. For example, the VGG19 CNN does almost as well with an $R^2_{model} = 46.84$. This and the robustness analyses (§4.5) suggest that the $R^2_{model} = 46.88$ from ResNet CNN is sufficient to demonstrate the value or prelaunch image-based features. Future research might explore custom deep-learning models or alternative pretrained models.

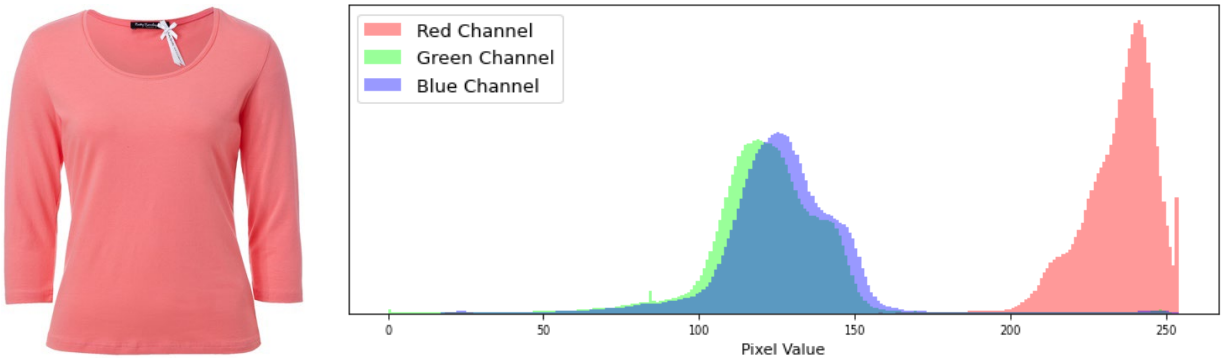
The results in Table 3 represent the performance of the model estimated on the entire data set including all product categories. We also explored per-category models for categories with sufficiently many sales and found that predictions in all five largest categories benefit from images, e.g., predictions in "Shirts" improves from $R^2_{model} = 14.78$ to 24.08 and predictions in "Dresses" improves from $R^2_{model} = 28.11$ to 31.13. All per-category models are in Online Appendix B. Machine-learning models are notoriously data hungry—the best predictions are obtained with a model that merges data from all categories.

4.5. Robustness of Incremental Predictive Ability Due to Image Features

Robustness of the basic hypothesis. Our basic hypothesis is that automated image-based features improve predictions and enable retailers to make more profitable display/not-display decisions. Table 3 is based on a particular set of image-based features (CNN) and a particular predictive model (GBRT). We have already summarized that for our data (1) the GBRT predicts best, but other machine-learning models are feasible, (2) alternative deep-learned image-based features are feasible, (3) the results are robust to evaluative criteria, (4) robust to precision weighting by the number of units purchased per item and (5) alternative data screening (minimum threshold of 10 or 30 rather than 20 items). The basic insights also hold for (6) dimensionality reduction of the 2,048 CNN features with various forms of principal component analysis (PCA) and (7) measures of uniqueness and distance from prior fashion seasons – details in Online Appendix B.

(Black box) automated pattern & color features. We test one more level of robustness. Researchers in machine learning often use automated pattern & color features as an alternative to deep-learning image-based features. Such color & pattern features might improve return-rate predictions. RGB color histograms provide one popular automated color feature. Figure 4 Illustrates an RGB coding of the color of an example fashion item as heavily based on red, but with mid-level peaks in green and blue. The number of bins in Figure 4, 256 x 256 x 256 $\approx$ 16 million, is too large for a GBRT. For feasibility we use 5 x 5 x 5 = 125 bins. To capture patterns, we use Gabor filters. Gabor filters use frequency-domain transforms to isolate the periodicity and the direction of that periodicity with sinusoidal waves (Manjunath and Ma 1996). Although Gabor filters are difficult to interpret, they might improve prediction. See Liu et al. (2020) for an application.

Figure 4. Example RGB Color Histogram Encoding of an Apparel Item



These automated pattern & color features do not predict as well as CNN-based features ($\boldsymbol{R^2_{model} = 45.28}$). Adding these features to a model based on CNN features and color labels is redundant (does not improve predictions, $\boldsymbol{R^2_{model} = 46.84}$). Alternative automated color features (HSV features, ORB features) do not change the basic message – details in Online Appendix B. Although automated pattern & color features provide an alternative to CNN features in the predictive model, they do not enhance interpretability. We examine more interpretable features in §5.

Results for human-coded features (HCF). The HCFs improve predictions relative to the non-image baseline (6.85% for dresses and shirts), but do not predict as well as the models with automated CNN-based features (9.92% dresses and shirts)[1]. Given the added time and cost of HCFs, the CNN features appear to be a better choice for the predictive model. The interpretable features in §5 are curated to generalize the HCFs. §5 suggests they predict better than the HCFs.

Summary of robustness tests. The GBRT/CNN model appears to be robust to alternative predictive models, alternative deep-learning image-processing features, alternative performance metrics, alternative data cleaning, dimensionality reduction, and the use of automated pattern & color features. The GBRT/CNN model appears to be a reasonable proof-of-concept. It is of course possible that

[1] We re-estimated all models for the two categories for which HCFs were coded (dresses and shirts). The absolute predictive ability, but not the relation among models, varies when we limit the data to the two largest categories.

some retailers will adopt alternative models or image-based features for reasons outside our analysis. The performance of many alternative models reinforces the basic hypothesis that image-based features help manage returns.

4.6. The Relationship between a Model's Predictive Accuracy and Profitability

Because all items are already inventoried for the bricks-and-mortar stores (§3.1), the marginal fixed costs for displaying the items online are minimal. As long as we do not compromise overall variety, we can consider removing (not displaying) items that have negative expected profits based on the predicted return rate, $\hat{r}_i^{model}$, and a measure of the uncertainty in $\hat{r}_i^{model}$. For the remainder of this section, we simplify notation and use $\hat{r}_i$ as shorter notation for $\hat{r}_i^{model}$. r_i continues to denote the true return rate.

We make online display/not-display decisions item by item. Because our estimate, $\hat{r}_i^{model}$, is independent of the number of items sold, $N_{i,purchased}$, and because fixed costs are negligible for displaying an item, we focus on profit per item sold.

This section evaluates whether we should not display items with negative expected profit per sale or whether we should take the precision of the model into account. For example, perhaps we should be more aggressive with a model that predicts better. On the other hand, if predictions were no better than random noise, perhaps we should be more cautious about not displaying any items.

There are inventory costs for carrying an item, but those are well-studied, present no new insight, and can easily be added to the profit-maximizing model. The firm's allowable-return policies are set by law and considered fixed for this analysis. To a first order, we ignore interactions among items and assume that the small percentage of items removed does not affect the demand for the remaining items. Fortunately, our the policy in our empirical application removes a small fraction of items. To the extent that removing some items increases demand for other items, the increased demand improves profits further. To the extent that removing some items decreases overall demand for the online store, profits

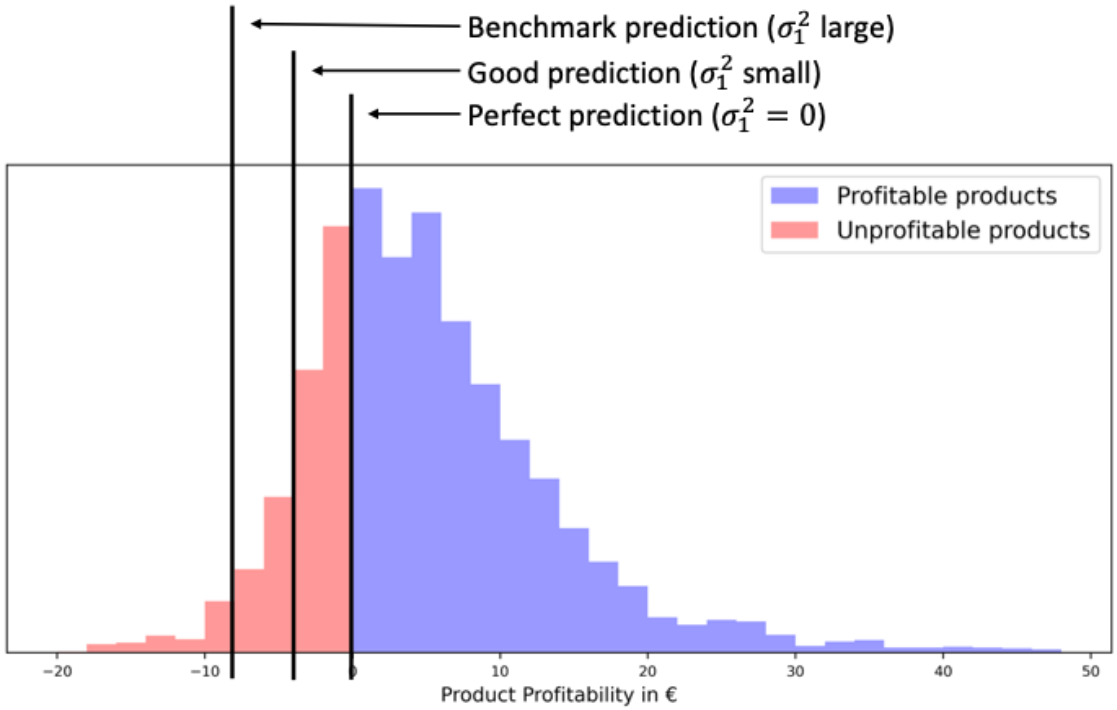
might decrease—an issue we do not have the data to address.

Naïve policy that ignores uncertainty in predictions. The return costs for returned items consist of two components: a flat processing cost for shipping and handling the return (c_{fix}) and a cost that is proportional to price of the item (c_{var}) because returned items must be discounted or discarded if they are damaged or out of season. Let p_i be the price of item i and c_i be the item's cost. If r_i were known, the profitability of item i, π_i, would be given by:

$$\pi_i = (1 - r_i)(p_i - c_i) - r_i(c_{fix} + p_i c_{var}). \quad (3)$$

The retailer's exact costs are proprietary. For illustration, we used a fixed return cost c_{fix} of 5.31€ (iBusiness 2016) and a variable return cost c_{var} of 13.1% of an item's price (Asdecker 2015). The naïve policy using these illustrative costs would imply that 27.2% of the items in our data are unprofitable. The naïve policy would remove these items. However, the predicted profit from the naïve policy is not achievable because our predictions are uncertain. The naïve policy would overestimate true profits as a 25% improvement. The naïve policy would also violate the assumption that a small percentage of items is not displayed. Figure 5 illustrates this naïve policy. "Profitable" products are retained (shown as blue in Figure 5) and the "unprofitable" products not displayed (shown as pink in Figure 5).

Figure 5. Distribution of Items' Profitability in Our Data



Policy taking uncertainty into account. Our empirical model produces imperfect predictions of the return rates, $\hat{r}_i$. We assume that $\hat{r}_i$ is unbiased. This implies that our estimate of profitability, $\hat{\pi}$, has a mean equal to the true profits with a variance based on the uncertainty in the predicted return rate. In symbols, $\hat{\pi}_i \mid \pi_i \sim \mathcal{N}(\pi;\ \sigma_1^2)$. We examine whether the retailer's decision depends on the ability of the model to accurately predict return rates, that is, we examine whether the policy depends on σ_1^2.

Because decisions are made for each item, i, we temporarily drop the item subscript, i. We assume the retailer's prior beliefs about profits are normally distributed across items: $\pi \sim \mathcal{N}(\mu_o;\ \sigma_o^2)$. Let $\mathcal{P}$ be a policy such that the retailer displays the item if $\mathcal{P} = 1$ and does not display the item if $\mathcal{P} = 0$. Let $\phi \equiv (\hat{\pi}, \mu_0, \sigma_0^2, \sigma_1^2)$, then the uncertainty-dependent policy is based on solving the following mathematical problem:

$$\max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}[\mathcal{P}(\phi) * \pi + \big(1 - \mathcal{P}(\phi)\big) * 0]. \tag{4}$$

The policy that maximizes the mathematical expression in Equation 4 is a threshold policy given

by Equation 5. Equation 5 yields intuitive policies as σ_o^2 and σ_1^2 approach zero (perfect information) or infinity (no information) and that expected profits using the policy decrease in σ_1^2.[2]

$$\mathcal{P}(\phi) = \begin{cases} 1 \; if \; \hat{\pi} \geq -\mu_o \frac{\sigma_1^2}{\sigma_o^2}, \\ 0 \; if \; \hat{\pi} < -\mu_o \frac{\sigma_1^2}{\sigma_o^2}. \end{cases} \tag{5}$$

Assuming that the retailer has positive priors, we added the thresholds for uncertainty-based policies to Figure 5. (1) For perfect predictions ($\sigma_1^2 = 0$), launch all items for which $\hat{\pi} > 0$. (2) For good predictions (σ_1^2 small), launch only items for which $\hat{\pi}$ exceeds the threshold. And (3), when predictions are extremely noisy (σ_1^2 large), launch almost all items even those with expected negative profits. As predictive uncertainty σ_i^2 increases, the uncertainty-based policy screens out fewer items and achievable profits decline. The dependency on σ_i^2 motivates MSE and R_{model}^2 as appropriate criteria with which to judge the predictive model. The better the R_{model}^2, the better is the achievable profit.

Using our data, we simulate the model-based policies (see Table 4). When the GBRT/CNN model is used to determine the data-based display/not-display policy, the retailer chooses not to launch 7.13% of the items. The expected profits increase by 8.29% relative to launching all the items. Even compared with the non-image baseline, the improvement in profits is important to fashion retailers with many items in many categories over many fashion seasons. This is especially true for fashion items that are high-priced and high-volume. The potential for profit improvement is even greater if retailers were able to source and/or improve items at the design stage. To that end, we next examine interpretable features, both image and non-image, that are associated with high and low return rates.

[2] Derivation of the threshold policy, a proof that expected profits decrease in σ_1^2, and limiting cases as σ_o^2 and σ_1^2 approach zero or infinity are provided in Online Appendix A.

Table 4. Expected Profit Improvement Using Different Predictive Models

Model	Features	Percent Items not Launched	Profit Improvement vs. Launch All Items
Non-image baseline	Category, seasonality, and price	5.98% (0.11)	6.81% (0.18)
Color labels added to baseline	Category, seasonality, price, and color labels	6.26% (0.13)	7.16% (0.19)
CNN Features	Category, seasonality, price, CNN from image	7.13% (0.12)	8.29% (0.23)

5. Generating Interpretable Insights

The retailer might improve its profits further if it were to use information available in images to make decisions when sourcing or designing new fashion items. To help the retailer's buyers source items and to help the retailer's designers design new items, we complement the predictive model with an interpretable model that identifies item features that are linked to high and low return rates. To deal with large assortments and rapid fashion seasons, we seek automatically-extracted image-based interpretable features that do not require consumer tests, surveys, or experiments. When the retailer can invest in HCFs, the HCFs enhance interpretability to the extent they help buyers and designers visualize the image-based features.

5.1. Automatically-extracted Image-based Interpretable Features

Each of the proposed image-based features is based on insights, experience, and expectations from the fashion industry. We seek features that are interpretable by the retailer's buyers and designers but can be generated at scale automatically. Automatic generation allows the retailer to use the features for large assortments in every fashion season. We extract features related to color (color clusters, color dominance, brightness, horizontal and vertical color asymmetry), pattern (pattern direction, pattern complexity), shape (shape asymmetry, shape ratio, shape triangularity), and item uniqueness (uniqueness). These features are chosen to be as general as feasible and, hopefully, subsume more-

specific image-based features such as the HCFs. We describe each of these features in detail below and provide examples in Figure 6. These features illustrate the type of automatically-generated features that are feasible. At the end of this section, we examine whether this set of features captures sufficient variation in return rates.

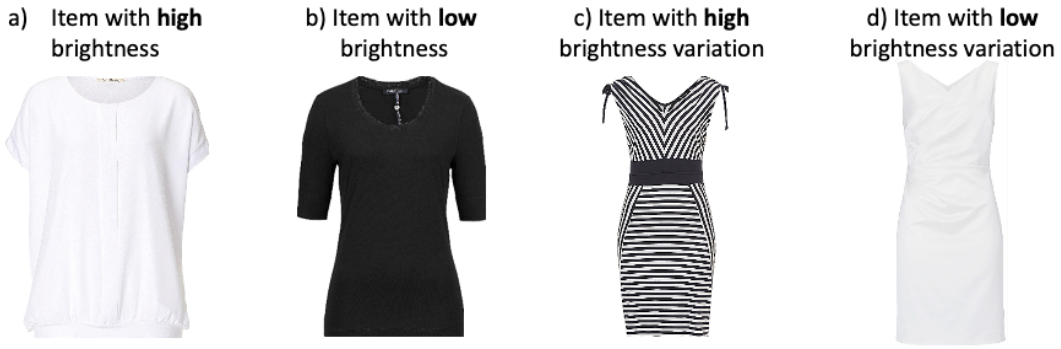
Color clusters. To visualize the basic color composition of an item, we use weighted K-means clusters in RGB-pixel space. For each item's image, we calculate the proportion of pixels closest to the mean of the color cluster (thirty clusters in our data). Unlike the retailer's color labels, the color clusters are more-nuanced and data-driven.

Color dominance. Some items have many colors but none dominate; other items have a dominant color with patterns, say flowers, of different colors. Color dominance is the maximum value of a color share for the item.

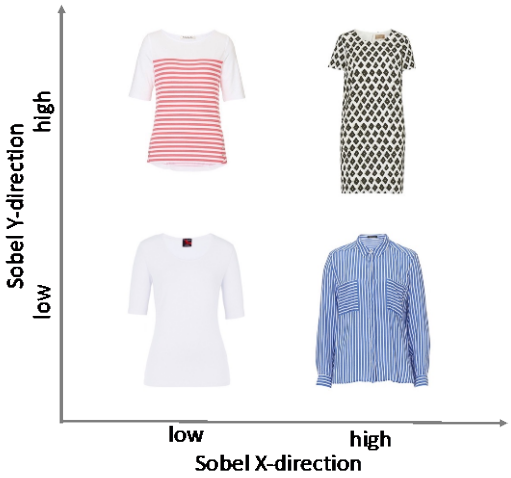
Brightness. The perceived brightness of an apparel item affects sales and return rates. Brightness might be partially redundant with color clusters, but that is an empirical question. **Brightness** is defined as the average intensity of the image after converting it to greyscale. Brightness varies over a garment. For example, if an item has a uniform color, the **brightness variation** is close to zero; if the item has a complex pattern of light and dark stripes, the brightness variation is larger. Computationally, we use the standard deviation. Both brightness and brightness variation are allowed to enter the model. Figure 6i illustrates fashion items with low and high brightness and brightness variation.

Figure 6. Examples of Automatically-extracted Image-based Interpretable Features

(i) Illustration of Items with High and Low Brightness and Brightness Variation



(ii) Illustration of Pattern Direction (Sobel X- and Y- Directions)



(iii) Illustration of Pattern Complexity

(iv) Illustration of Shape Ratio

Pattern direction. Gabor pattern features had moderate success in predicting returns (§4.5), but they are very difficult to interpret. Pattern direction and pattern complexity are more interpretable. Pattern direction is summarized by applying a Sobel filter to each direction (X for horizontal and Y for vertical) to the greyscale images (Gonzalez and Woods 2018). This is equivalent to a partial derivative with respect to movement orthogonally along either the horizontal or vertical axis. For example, **horizontal stripes** have a high derivative in the vertical direction and **vertical stripes** a high derivative in the horizontal direction (see Figure 6ii).

Pattern complexity. Some apparel items have checkered patterns (high derivative in both the horizontal and vertical directions), while others have more complex patterns. To represent pattern complexity, we extract edges from the image using the Canny edge detector and we extract straight lines using Hough transformations (Duda and Hart 1972). Each line is represented by the orthogonal distance from the top left corner of the image to the line and by the angle of the line relative to the X-axis. Two features are extracted: **pattern complexity** is the standard deviation of the angles of the extracted lines; the other feature is the **number of extracted lines.** Pattern complexity is extracted if there are more than twenty lines, otherwise it is set to zero. All such meta-parameters are tuned. Figure 6iii illustrates (a) an item with high pattern complexity (lines of varying angles) and (b) an item with low pattern complexity (horizontal stripes with a zero angle).

Asymmetry. The HCF analysis suggests that asymmetric items have higher return rates (review Figure 3). Shape asymmetry, horizontal color asymmetry, and vertical color asymmetry are likely to affect return rates. Dresses, shirts, and other apparel items are naturally asymmetric vertically. To extract **shape asymmetry**, we compare the left half of the image to the mirror image of the right half of the image. The percentage of non-overlapping pixels indicates shape asymmetry. For example, if the item is perfectly symmetric horizontally, then there will be no non-overlapping pixels; if the fashion item is highly asymmetric, there will be many non-overlapping pixels. To extract **horizontal color asymmetry**,

we use KL-divergence to compare the RGB histograms for the right and left halves of the image. **Vertical color asymmetry** compares the top and bottom halves.

Geometric shape. In fashion, the shape of an item is likely to be important in predicting return rates. For example, long dresses are often bought for more formal wear where fashion fit might be extremely important, while shorter dresses are bought for more casual wear where the consumer is less discerning. **Shape ratio**, the ratio of median width to the median height, captures both sleeveless and item-length phenomena. Because the GBRT allows interactions between the shape ratio and category, the impact of this variable can vary by category such as dresses (length matters more) versus shirts (sleeves matter more). Figure 6iv provides examples of high and low shape ratios for dresses and shirts. **Shape triangularity**, the ratio of the median width of the bottom 25% of the item to the median width of the top 25% of the item, differentiates many fashion items. For example, an A-line dress has high shape triangularity while a pencil dress has a shape triangularity close to 1. Because triangularity is easy to visualize, for brevity we do not provide examples in Figure 6.

Uniqueness. Uniqueness might contain information not otherwise captured by the automatically-extracted image-based interpretable features. For consistency with the GBRT/CNN model, we define uniqueness as the Euclidean distance between the CNN-learned features of the item and the category mean of the CNN-learned features. Although uniqueness did not improve the GBRT/CNN predictive ability, the lack of improvement may have been because the CNN features already contain a (black-box) measure that captures uniqueness.

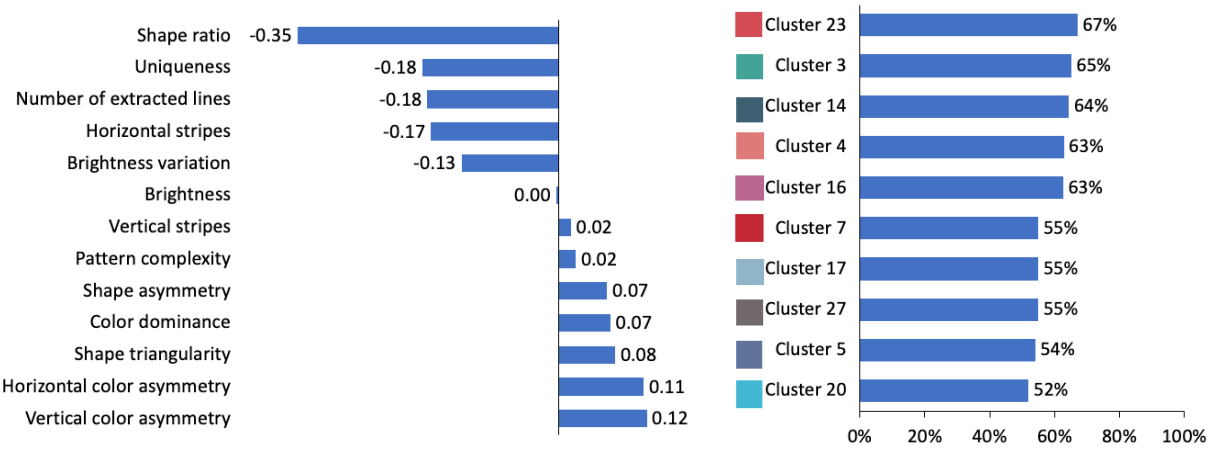
5.2. Model-Free Evidence Motivates the Use of Interpretable Features

Before we examine formal models, we examine whether or not the proposed automatically-extracted image-based interpretable features are related to return rates. Figure 7a reports the correlations between return rates and the interpretable features (other than color clusters). Figure 7b reports the return rates for the top five and bottom five color clusters. These model-free analyses

motivate a more-complex machine-learning model.

Figure 7. Model-Free Evidence: Correlations of Return Rates and Interpretable Image-based Features

(a) Correlations of Return Rate and features (b) Return Rate by Selected Color Clusters



Note: For illustration, we use in (b) the largest color cluster present in the product image.

5.3. Summarizing the Marginal Effect of the Interpretable Features

Our analyses in §4 suggest that features (image and non-image) interact and that a GBRT (or another machine-learning model) is a good model with which to predict return rates. For comparison to the CNN-based predictive model, we estimate a GBRT model with automatically-extracted image-based interpretable features added to non-image features. Interpreting the impact of features in a tree-based model with hundreds of trees is challenging. The machine-learning literature uses the SHAP (SHapley Additive exPlanations) framework to interpret the marginal impact of each feature on the predicted target variable (Lundberg & Lee 2017). The SHAP value is based on Shapley values from game theory and enables us to interpret feature impacts in an arbitrary black-box model. The SHAP value for feature j for item i (denoted as ϕ_{ij}) indicates the marginal change in the predicted return rate $\hat{r}_i$ due to a change in the value of interpretable feature j while taking into account all other features in the model. Mathematically, the predicted value $\hat{r}_i$ of the model for item i could be decomposed as $\hat{r}_i = \overline{r} + \sum_j \phi_{ij}$ (where $\overline{r}$ is the average return rate for all items). By computing SHAP values for all items, we obtain a

sample of SHAP values that can interpreted as the marginal impact on predicted $\hat{r}_i$ given a random set of all feature values.

To determine the relative importance of each interpretable image feature, we use the mean absolute SHAP value, $F_j = I^{-1} \sum_i \left| \phi_{ij} \right|$, where I is the number of items and we sum ϕ_{ij} over all items i for feature j. Intuitively, F_j measures how far on average the given feature j pushes the predicted value of $\hat{r}_i$ from the sample mean $\overline{r}$. For ease of interpretation, we use Pareto charts that rank the features by F_j and display the most impactful features first.

To illustrate the use of SHAP values, Figure 8 ranks the color cluster centers by their impact on the predicted return rate, F_j, from highest to lowest, measured by the average absolute SHAP value within the cluster. Figure 8 provides more nuanced interpretations on an item's color composition than do retailer pre-defined color labels. For example, prototypical red has a high impact, but other shades of red have a low impact. Some shades of blues have a high impact, but the prototypical blue has a low impact.

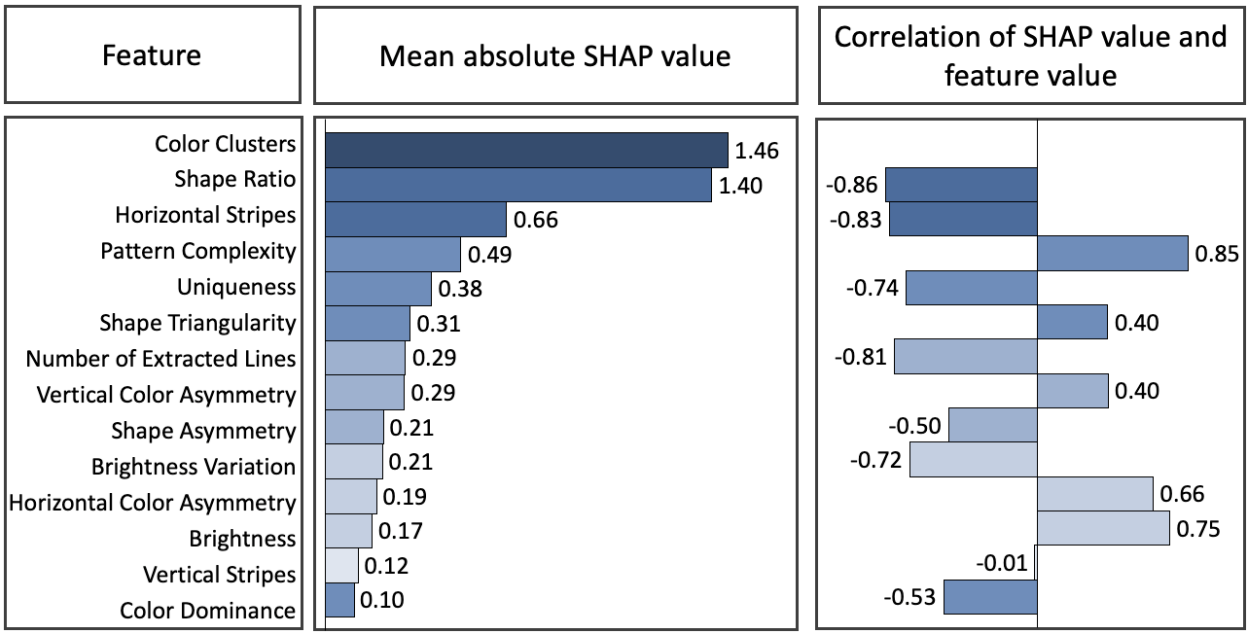
Figure 8. The Effect of Color Clusters on Return Rates (Clusters Ranked by F_j)

5.4. Automatically-Extracted Image Features and Return Rates

Figure 9 provides the Pareto Chart for the automatically-extracted image-based interpretable features. We address non-image features in §5.6. Figure 9 reports mean-absolute SHAP values for all features. We aggregate the impact of color clusters by the sum of their SHAP values. Online Appendix B provides more detail on color clusters

The F_j do not indicate the direction of the impact of a feature, nor do the F_j illustrate variation across items. For example, a feature may have the same F_j value if it is high on a few items and low on many, or just moderate for all items. Figure 9 complements F_j with the correlation between SHAP values and standardized feature values to indicate the direction of impact on predicted return rate and to suggest whether the feature affects many items (high correlation) or just a few items (low correlation). See Online Appendix B for the directionality and variation of impact of color clusters.

Figure 9. Pareto Chart of SHAP Values for the Automatically-Extracted Image Features



Consistent with experience in fashion apparel, color clusters have the greatest impact on return rates. Shape ratios are the next most important and the direction is as expected. More formal items (lower shape ratio) have higher return rates than more casual items (higher shape ratio). As expected,

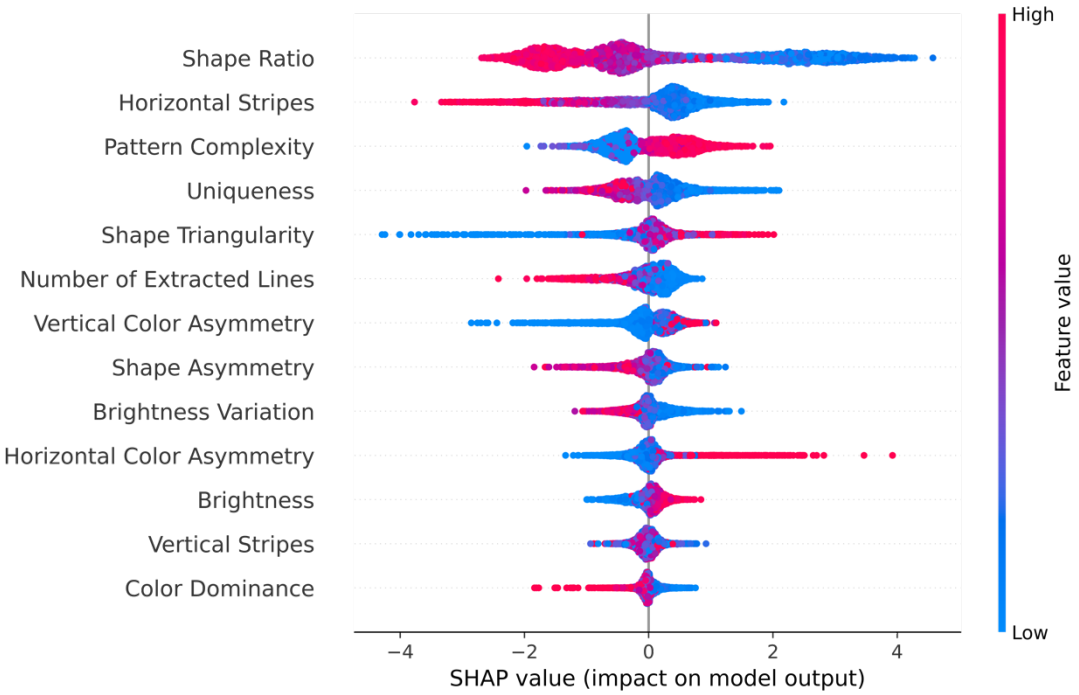
shape ratio has different interpretations for different categories (captured in Figure 10 below). Sleeveless dresses (lower shape ratio) are returned more often, consistent with the implications of the HCFs (see §3.2). Interestingly, horizontal stripes are important and are associated with low return rates, while vertical stripes are much less important. Pattern complexity is important and positively correlated with return rates, while uniqueness is negatively correlated with return rates. Uniqueness was redundant with the CNN features in the predictive model (§4.5), but provides incremental predictive ability in a model with interpretable features (Figure 9).

The brightness features are less important, likely because some brightness information is extracted by the color clusters. However, bright products have higher return rates while products with a high variation of brightness (many contrasting colors) have lower return rates. Interestingly, brightness has low impact on predicted return rate, but high correlation. When we examine variation among items (Figure 10), this is explained because the SHAP values tend to be small in magnitude, but consistent in their impact on return rates.

To provide further insight to buyers and designers about the variation in SHAP values within items, we use a method known as "bee-swarm charts" to visualize the SHAP values, ϕ_{ij} for all items i for all features j. A bee-swarm chart details, for each item, the impact on return rate predictions of high vs. low values of the feature. Features are ranked by the mean absolute SHAP values, F_j.

Figure 10 provides the variation in impact (bee-swarm chart) for the automatically-extracted interpretable image-based features. For example, on average items with higher values of the shape ratio are less likely to be returned, but this relation is not homogeneous. Buyers and designers can examine the detailed points, each of which corresponds to an item, to determine the impact of that item's shape ratio for that item. This can be done for any of the automatically-extracted interpretable image-based features, including color clusters. (Online Appendix B provides bee-swarm charts for color clusters.)

Figure 10. The Impact of Automatically-Extracted Interpretable Image-based Features (Bee-swarm chart)



5.5. Using the Interpretable Features to Source or Design Fashion Items

By combining the insights from Figures 6 through 10, we can predict items that are likely or not likely to be returned. For example, shirts with higher shape ratios, horizontal stripes, and darker colors (red dots to the left in the bee-swarm chart) are less likely to be returned. Shirts with lower shape ratios, solid colors (no horizontal stripes or patterns), and a pinkish color (cluster 24) are more likely to be returned (blue dots to the right in the bee-swarm chart). Examples of such shirts and dresses are shown in Figure 11. For confidentiality, we do not provide the predicted or actual return rates, but they are consistent with the expectations from the interpretable model.

Figure 11. Illustration of Combining Interpretable Image Features on Expected Return Rates

Note: We expect that Shirt (a) and Dress (a) have low return rates and Shirt (b) and Dress (b) to have high return rates. The data confirms these expectations.

5.6. Non-image Features and Return Rates

Item category, price, and seasonality are all important features. With mean absolute SHAP values of 2.89, 2.78, and 1.37, respectively, they are, on average, more impactful than the automatically-generated image-based interpretable features. Non-image features are best included in the GBRT model from which SHAP values are computed, both as controls and because of their interactions with the automatically-extracted image-based interpretable features. The non-image features also provide valuable diagnostic information. Online Appendix B provides greater detail on the non-image features, e.g., sales and return rates by category.

5.7. Comparison of Predictive Models: Deep-learned vs. Interpretable Features

Recognizing the tradeoff between predictive ability and interpretability, we expect an interpretable-feature GBRT model to predict better than either the non-image baseline or the color-label model, but we do not expect the model to predict as well as the GBRT/CNN model. This is indeed the case: a GBRT model based on the automatically-extracted image-based interpretable features has an $R^2_{model} = 45.81$, which is less than that the $R^2_{model} = 46.88$ for the GBRT/CNN model. Both predictive

abilities are well above the non-image baseline and the rudimentary image (color-label) model. Predictive ability is slightly better than the more-difficult-to-interpret automated-pattern-&-color-features model (Table 1) and better than the model based on HCFs (§4.5).

Our automatically-generated image-based interpretable features were curated carefully to provide insight while predicting return rates, but such choices are not unique. Retailers and researchers may wish to explore other interpretable features or combinations of HCFs and interpretable features.

6. Discussion and Further Research

Product returns generate considerable costs for online retailers – a large and growing retail channel. We propose that images, available prior to a fashion season, enable retailers to select which fashion-items to display online. We demonstrate, by example, that image-based features in a machine-learning model provide substantial incremental predictive ability relative to models based on traditional measures available to the retailer prior to launch. The predictive ability appears to be robust to a large number of variations. The display policy depends on the accuracy of the predictions and demonstrates that increased profits are feasible.

We augment predictions with automatically-extracted image-based interpretable features that can be used quickly and repeatedly for every fashion season and that scale to large assortments and many categories of items. The interpretable model sacrifices a small amount of predictive ability to provide diagnostic information valuable to the retailer's buyers and designers. Both the predictive and interpretable models, once developed and trained, run quickly and scale well.

Our application focuses on fashion-item returns in the apparel industry. This industry is important by itself, but we expect the approach to apply more broadly. Incorporating product images has the potential to improve predictive accuracy prior to product launch and generate important insights for design in industries such as hospitality, furniture, real estate, and even groceries.

Our analyses are illustrative and *ceteris paribus*. Researchers might explore (1) policies in which

items are displayed online but not offline, (2) the implications on overall demand (online and offline) of not displaying items online, (3) interactions among items, (4) policies in which online items can be returned offline and thus increase offline traffic, (5) analyses that combine prelaunch features with postlaunch features, (6) how item features and prices jointly affect return rates, and (7) models that predict and provide insight jointly about sales and returns.

Appendix. Tuning of Hyperparameters of the GBRT Model

We tuned the hyperparameters of the GBRT model with a grid search over the set of parameters presented in Table C.1. The criterion was predictive ability in the validation sample. The model was then tested using the held-out out-of-sample predictions. For each iteration of the grid search, we stopped adding additional regression trees after the accuracy on the validation sample did not improve for twenty-five consecutive trees.

Table C.1. Grid for the GBRT Hyperparameters

LightGBM parameter name	Set of tested values	Parameter Description
n_estimators	$[3000]$	Maximum number of boosting trees
learning_rate	$[0.025, 0.01, 0.05]$	Shrinkage rate
max_depth	$[7, 9, 11]$	Maximum depths of the regression tree
num_leaves	$[32, 48]$	Maximum number of leaves in one regression tree
reg_lambda	$[0, 5]$	Weight of L2 regularization
reg_alpha	$[0, 5]$	Weight of L1 regularization
colsample_bytree	$[0.5]$	Random subset of features to be used in one regression tree

Note: Parameters not listed in the table take default values in LightGBM package

References

Anderson ET, Hansen K, Simester D (2009) The option value of returns: Theory and empirical evidence. *Marketing Science. 28*(3): 405**–**423.

Asdecker B (2015) Returning mail-order goods: analyzing the relationship between the rate of returns and the associated costs. *Logistics Research*. 8(1):1**–**12.

Bower AB and Maxham-III JG (2012) Return shipping policies of online retailers: Normative assumptions and the long-term consequences of fee and free returns. *Journal of Marketing*, 76(5):110–124.

Conlin M, O'Donoghue T, Vogelsang T. (2007) Projection bias in catalog orders. *American Economic Review*. 97(4):1217–1249.

Davis S, Hagerty M, Gerstner E (1998) Return policies and the optimal level of "hassle". *Journal of Economics and Business*. 50 (2): 445-460

Dew R, Ansari A, Toubia O (2022) Letting Logos Speak: Leveraging Multiview Representation Learning for Data-Driven Logo Design. *Marketing Science*. Forthcoming.

Duda R, Hart P (1972) Use of the Hough transformation to detect lines and curves in pictures. *Communications of ACM*. 15 (1): 11-15

Dzyabura D, El Kihal S, Peres, R (2021) Image analytics in marketing. In: Homburg, C, Klarmann, M, Vomberg, AE (eds) Handbook of Market Research. Springer, Cham.

El Kihal S, Nurullayev N, Schulze C, Skiera B (2021), A comparison of product return rate calculation methods: evidence from 16 retailers. *Journal of Retailing*. 97(4): 676–696.

El Kihal S, Shehu E (2022) It's not only what they buy, it's also what they keep: Linking marketing instruments to product returns. *Journal of Retailing*. Forthcoming.

Gonzalez R, Woods R (2018) Digital Image Processing (4th edition). MA: Addison-Wesley.

Hartmann J, Heitmann M, Schamp C, Netzer, O (2021) The Power of brand selfies. *Journal of Marketing Research.* 58(6): 1159-1177.

Hauser, J. (1978). Testing the accuracy, usefulness, and significance of probabilistic choice models: An information-theoretic approach. *Operations Research.* 26(3), 406-421.

He R, McAuley J (2016) VBPR: Visual Bayesian personalized ranking from implicit feedback. *Proceedings* of the Thirtieth AAAI Conference on Artificial Intelligence.

He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. arXiv preprint https://arxiv.org/abs/1512.03385.

Hernández-Orallo J (2013) ROC curves for regression. *Pattern Recognition.* 46(12):3395-3411.

Hong Y, Pavlou PA (2014) Product fit uncertainty in online markets: Nature, effects, and antecedents. Information Systems Research. 25(2): 328-344.

iBusiness (2016) Wie Shopbetreiber das Retourenproblem wirklich lösen [How Online Retailers are Solving the Problem with Product Returns], http://www.ibusiness.de/aktuell/db/858729veg.html

Janakiraman N, Syrdal HA, Freling R (2016) The effect of return policy leniency on consumer purchase and return decisions: A meta-analytic Review. Journal of Retailing. 92(2):226–235.

Klostermann J, Plumeyer A, Böger D, Decker R (2018) Extracting brand information from social networks: Integrating image, text, and social tagging data. *International Journal of Research in Marketing*, 35(4): 538–556.

Liu J, Toubia O (2018) A semantic approach for estimating consumer content preferences from online search queries. Marketing Science. 37(6):930–952.

Liu L, Dzyabura D, Mizik NV (2020) Visual listening in: Extracting brand image portrayed on social media. Marketing Science. 39(4):669–686.

Lundberg S, Lee S (2017) A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems. 4768-4777

Lynch C, Aryafar K, Attenberg J (2015) Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank. arXiv preprint arXiv:1511.06746.

Manjunath BS, Ma WY (1996) Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 18(8): 837–842.

McAuley J, Targett C, Shi Q, van den Hengel A (2015) Image-based recommendations on styles and substitutes. arXiv preprint: arXiv:1506.04757

Moorthy S and Srinivasan K (1995) Signaling quality with a money-back guarantee: The role of transaction costs. *Marketing Science*, 14(4):442–466.

Narang U, Shankar V (2019) Mobile app introduction and online and offline purchases and product returns. *Marketing Science*. 38(5): 756-772

Petersen J and Kumar V (2009) Are product returns a necessary evil? Antecedents and consequences. *Journal of Marketing*, 73(3):35–51.

Petersen J and Kumar V (2010) Can product returns make you money? *MIT Sloan Management Review.* 51(3):85.

Petersen J and Kumar V (2015) Perceived risk, product returns, and optimal resource allocation: evidence from a field experiment. *Journal of Marketing Research.*52(2):268-285.

Sahoo N, Dellarocas C, Srinivasan S (2018) The impact of online product reviews on product returns. *Information System Research*. 29 (3):525-777.

Shehu E, Papies D, Neslin S (2020). Free shipping promotions and product returns. *Journal of Marketing Research*. 57(4): 640-658.

Shi M, Chussid C, Yang P, Jia M, Lewis VD, Cao W (2021) The exploration of artificial intelligence application in fashion trend forecasting. *Textile Research Journal* 91(19-20):2357-2386.

Shulman JD, Coughlan AT, Savaskan RC (2011) Managing consumer returns in a competitive environment. *Marketing Science.* 57(2): 347–362.

Thomasson E (2013) Online retailers go Hi-Tech to size up shoppers and cut returns. *Reuters* (October 2), http://www.reuters.com/article/net-us-retail-online-returns-idUSBRE98Q0GS20131002

Wall Street Journal (2022) Retailers' many unhappy returns. https://www.wsj.com/articles/retailers-many-unhappy-returns-11641387605

Wood S (2001) Remote purchase environments: the influence of return policy leniency on two-stage decision processes. *Journal of Marketing Research.*38(2):157-169.

Zhang M, Luo L (2022) Can user generated content predict restaurant survival: Deep learning of Yelp photos and reviews. *Management Science*. 68(8): 5644-5666.

Zhang S, Lee D, Singh PV, Srinivasan K (2021) How much is an image worth? The impact of professional versus amateur Airbnb property images on property demand. *Management Science*. Forthcoming.

Online Appendix A. Proof that the Profit-Maximizing Policy is a Threshold Policy, that the Threshold Policy is Intuitive, and that Profits are Decreasing in Predictive Uncertainty.

Result 1. Suppose (1) the firm's prior on the profitability of an item, π, is normally distributed, $\pi \sim \mathcal{N}(\mu_0;\ \sigma_0^2)$, (2) the firm observes an estimate of profitability $\hat{\pi}|\pi \sim \mathcal{N}(\pi;\ \sigma_1^2)$, and (3) the firm seeks a policy to decide whether to put an item online or not. Then the profit maximizing policy, $\mathcal{P}(\phi)$, is a threshold policy:

(A1) $$\mathcal{P}(\phi) = \begin{cases} 1\ if\ \hat{\pi} \geq -\mu_o \frac{\sigma_1^2}{\sigma_o^2} \\ 0\ if\ \hat{\pi} < -\mu_o \frac{\sigma_1^2}{\sigma_o^2} \end{cases}$$

The policy in Equation A1 is intuitive. For example,

- If predictions are perfect, then $\sigma_1^2 = 0$ and the policy reverts to that of perfect prediction; launch those items for which $\hat{\pi}\ \geq 0$.
- If the model has no predictive ability, then $\sigma_1^2 \to \infty$ and the policy reverts to the prior mean, μ_o; launch all items if and only if the prior mean is positive.
- If there is no uncertainty in the prior, then $\sigma_o^2 \to 0$ and the policy again reverts to the prior mean; launch all items if and only if the prior mean is positive.
- For finite values of σ_1^2 and σ_o^2, the ratio, σ_1^2/σ_o^2, modifies the amount by which the predicted profits must exceed prior beliefs in order to launch.

***Proof of the threshold policy**:* The firm solves the following optimization problem:

(A2) $$\max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}[\mathcal{P}(\phi) * \pi + \big(1 - \mathcal{P}(\phi)\big) * 0] = \max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}[\mathcal{P}(\phi) * \pi]$$

where $\phi \equiv (\hat{\pi}, \mu_0, \sigma_0^2, \sigma_1^2)$ is the set of all known parameters; $\hat{\pi}|\pi \sim \mathcal{N}(\pi;\ \sigma_1^2)$ and $\pi \sim \mathcal{N}(\mu_0;\ \sigma_0^2)$

Using the law of iterative expectations, we rewrite the initial maximization problem (A2) as:

(A3) $$\max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}[\mathcal{P}(\phi) * \pi] = \max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}\big[\mathcal{P}(\phi) * \mathbb{E}[\pi|\phi]\big] = \max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}\big[\mathcal{P}(\phi) * \mathbb{E}[\pi|\hat{\pi}]\big]$$

The last step relies on the assumption that $\sigma_0, \sigma_1, \mu_0$ are observable.

Because $\mathbb{E}[\pi|\phi]$ is a function of observables, ϕ, we can denote $\mathbb{E}[\pi|\phi] = f(\phi)$. Equation (A3) is rewritten as:

(A4) $$\max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}[\mathcal{P}(\phi) * f(\phi)]$$

Equation (A4) implies that the optimal policy $\mathcal{P}^*(\phi)$ has the following form ($\mathcal{I}(\cdot)$ is an indicator function):

(A5) $$\mathcal{P}^*(\phi) = \mathcal{I}(f(\phi) \geq 0) = \mathcal{I}(\mathbb{E}[\pi|\phi]) \geq 0)$$

We show in the following that, for the case of normal priors, this policy would have a threshold form. [Note that the optimal policy in Equation (A5) does not depend on the normality assumption profitability; the policy is easily generalized to other distributions.]

Because $\hat{\pi}$ is normally distributed conditionally on π and since the prior is also normally distributed, the posterior is normally distributed. Using standard formulae, we write:

(A6) $$\pi|\hat{\pi}\sim\mathcal{N}\left(\frac{\hat{\pi}\sigma_0^2+\mu_0\sigma_1^2}{\sigma_0^2+\sigma_1^2}; \frac{\sigma_0^2\sigma_1^2}{\sigma_0^2+\sigma_1^2}\right) \quad \text{and} \quad \hat{\pi}\sim\mathcal{N}(\mu_0; \sigma_0^2+\sigma_1^2)$$

From (A6), it follows that:

(A7) $$\mathbb{E}[\pi|\phi] = \frac{\hat{\pi}\sigma_0^2+\mu_0\sigma_1^2}{\sigma_0^2+\sigma_1^2} \Rightarrow \mathcal{P}^*(\phi) = \mathcal{I}(\mathbb{E}[\pi|\phi] \geq 0) = \mathcal{I}\left(\hat{\pi} \geq -\mu_0 * \frac{\sigma_1^2}{\sigma_0^2}\right)$$

Which is the threshold policy.

Result 2. Under the assumptions of Result 1, the optimal expected profit is:

(A8) $$\Pi^* = \left(1 - \Phi\left(-\frac{\mu_0}{\sigma_\nu}\right)\right) * \mu_0 + \sigma_\nu * \varphi\left(-\frac{\mu_0}{\sigma_\nu}\right)$$

Where $\Phi(\cdot)$ and $\varphi(\cdot)$ are the standard normal CDF and PDF respectively, and $\sigma_\nu = \frac{\sigma_0^2}{\sqrt{\sigma_0^2+\sigma_1^2}}$.

***Proof**: B*y substituting the optimal policy from (A7) and conditional expectation from (A8) to (A2), we rewrite the expected optimal profit as:

(A9) $$\Pi^* = \mathbb{E}\left[\mathcal{I}\left(\hat{\pi} \geq -\mu_0 * \frac{\sigma_1^2}{\sigma_0^2}\right) * \left(\frac{\hat{\pi}\sigma_0^2+\mu_0\sigma_1^2}{\sigma_0^2+\sigma_1^2}\right)\right] = \mathbb{E}[\mathcal{I}(\nu \geq 0) * \nu] = \mathbb{P}[\nu \geq 0]\mathbb{E}[\nu|\nu \geq 0]$$

where $\nu = \frac{\hat{\pi}\sigma_0^2+\mu_0\sigma_1^2}{\sigma_0^2+\sigma_1^2} \sim \mathcal{N}\left(\frac{\hat{\pi}\sigma_0^2+\mu_0\sigma_1^2}{\sigma_0^2+\sigma_1^2}; \frac{\sigma_0^4}{(\sigma_0^2+\sigma_1^2)^2}(\sigma_0^2+\sigma_1^2)\right) \sim \mathcal{N}\left(\mu_0; \frac{\sigma_0^4}{\sigma_0^2+\sigma_1^2}\right) \sim \mathcal{N}(\mu_0; \sigma_\nu^2)$

Because ν is normally distributed, (A9) can be rewritten using the formula for the expectation of the truncated normal distribution:

(A10) $$\Pi^* = \left(1 - \Phi\left(-\frac{\mu_0}{\sigma_\nu}\right)\right) * \mu_0 + \sigma_\nu * \varphi\left(-\frac{\mu_0}{\sigma_\nu}\right)$$

Result 3. The expected profit under the optimal policy is a decreasing function of σ_1^2.

Proof: Taking the derivative of (A10) with respect to σ_1^2:

(A11) $$-\mu_0 * \varphi\left(-\frac{\mu_0}{\sigma_\nu}\right)\left(-\frac{\mu_0}{2\sigma_0^2(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}}\right) - \frac{\sigma_0^2}{2(\sigma_0^2+\sigma_1^2)^{\frac{3}{2}}}\varphi\left(-\frac{\mu_0}{\sigma_\nu}\right) +$$

$$\frac{\sigma_0^2}{(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}}\varphi'\left(-\frac{\mu_0}{\sigma_\nu}\right)\left(-\frac{\mu_0}{2\sigma_0^2(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}}\right) = \left(\frac{\mu_0^2}{2\sigma_0^2(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}} - \frac{\sigma_0^2}{2(\sigma_0^2+\sigma_1^2)^{\frac{3}{2}}} + \right.$$

$$\left.\frac{\sigma_0^2}{(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}}\left(\frac{\mu_0(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}}{\sigma_0^2}\right)\left(-\frac{\mu_0}{2\sigma_0^2(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}}\right)\right)\varphi\left(-\frac{\mu_0}{\sigma_\nu}\right) = \left(\frac{\mu_0^2}{2\sigma_0^2(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}} - \frac{\sigma_0^2}{2(\sigma_0^2+\sigma_1^2)^{\frac{3}{2}}} + \right.$$

$$\left.\left(-\frac{\mu_0^2}{2\sigma_0^2(\sigma_0^2+\sigma_1^2)^{\frac{1}{2}}}\right)\right)\varphi\left(-\frac{\mu_0}{\sigma_\nu}\right) = -\frac{\sigma_0^2}{2(\sigma_0^2+\sigma_1^2)^{\frac{3}{2}}}\varphi\left(-\frac{\mu_0}{\sigma_\nu}\right)$$

Because $\varphi(\cdot) > 0$ and $-\frac{\sigma_0^2}{2(\sigma_0^2+\sigma_1^2)^{\frac{3}{2}}} < 0$, the expected profitability is decreasing function of σ_1^2 and therefore an increasing function of model accuracy.

Online Appendix B. Supporting Tables and Figures

Table B.1. Improvement in Predictive Accuracy Varying Minimum Threshold on Online Sales

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over baseline
CNN Features with **10** as threshold for online sales	Category, seasonality, price, color labels	Deep-learning	43.14 (0.20)	+12.75%
CNN Features with **20** as threshold for online sales	Category, seasonality, price, color labels	Deep-learning	46.88 (0.19)	+13.48%
CNN Features with **30** as threshold for online sales	Category, seasonality, price, color labels	Deep-learning	51.23 (0.17)	+12.08%

Note: Improvements are calculated for baseline models estimated on the corresponding samples. Standard deviations are reported in parentheses.

Table B.2. Tests of Uniqueness, Precision (variance of $N_{i,purchased}$), and Distance from Prior Collections

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over CNN Features Model
CNN Features	Category, seasonality, price, color labels	Deep-learning	46.88 (0.19)	0.00%
CNN Features (including uniqueness)	Category, seasonality, price, color labels, image uniqueness	Deep-learning	46.82 (0.23)	-0.13%
CNN Features (including vs. last year)	Category, seasonality, price, color labels, image uniqueness	Deep-learning	44.55 (0.39)	-0.60% (see note)
CNN Features (precision weighting)	Category, seasonality, price, color labels, variance weighting	Deep-learning	46.77 (0.26)	-0.23%

Notes: The sample of items included when estimating the last-year model exclude products sold only in the first year of the data. A GBRT/CNN model for the same items yields 44.85 (0.33). The –2.8% is relative to this model. Standard deviations are reported in parentheses.

Table B.3. Improvement in Predictive Accuracy Using Alternative Prediction Models

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over Baseline
GBRT (CNN Features)	Category, seasonality, price, color labels	Deep-learning	46.88 (0.19)	+13.48%
Bagging Methods (CNN Features)	Category, seasonality, price, color labels	Deep-learning	45.35 (0.14)	+9.78%
LASSO (CNN Features)	Category, seasonality, price, color labels	Deep-learning	44.11 (0.32)	+6.78%

Note: Standard deviations are reported in parentheses.

Table B.4. Predictions for the two Largest Categories (Dresses and Shirts)

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over Baseline
Non-Image Baseline	Category, seasonality, price	None	57.94 (0.27)	–
Color-labels	Category, seasonality, price, and color labels	None	59.79 (0.24)	+3.19%
Automated Color Features	Category, seasonality, price, color labels	RGB	60.86 (0.22)	+5.04%
Automated Color and Patterns	Category, seasonality, price, color labels	RGB + Gabor	61.91 (0.31)	+6.85%
Human-coded features	Category, seasonality, price, color labels, human-coded features	Human-coded	61.91 (0.27)	+6.85%
CNN Features	Category, seasonality, price, color labels	Deep-learning	63.69 (0.19)	+9.92%

Note: Standard deviations are reported in parentheses.

Table B.5. Improvement in Predictive Accuracy Using an Alternative CNN

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over Baseline
ResNet CNN (this paper)	Category, seasonality, price, color labels	Deep-learning	46.88 (0.19)	+13.48%
VGG-19 CNN	Category, seasonality, price, color labels	Deep-learning	46.84 (0.18)	+13.39%

Note: Standard deviations are reported in parentheses.

Table B.6. Improvement in Predictive Accuracy Using PCA (nonlinear and linear tested; linear shown)

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over Baseline
Color Features	Category, seasonality, price, color labels	RGB	43.48 (0.18)	+5.25%
Color and Patterns	Category, seasonality, price, color labels	Gabor	41.37 (0.33)	+0.15%
CNN Features	Category, seasonality, price, color labels	Deep-learning	46.55 (0.21)	+12.68%

Note: Standard deviations are reported in parentheses.

Table B.7. Predictions Using Automated Pattern & Color Image-Processing Features

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over Baseline
Non-Image Baseline	Category, seasonality, price	None	41.31 (0.18)	–
Color Features	Category, seasonality, price, color labels	RGB	44.06 (0.20)	+6.66%
Pattern Features	Category, seasonality, price, color labels	Gabor	44.34 (0.23)	+7.33%
Color and Patterns	Category, seasonality, price, color labels	RGB + Gabor	45.28 (0.18)	+9.61%
CNN Features	Category, seasonality, price, color labels	Deep-learning	46.88 (0.19)	+13.48%
CNN Features all images	Category, seasonality, price, color labels	Deep-learning	47.48 (0.22)	+14.93%

Notes: All models use LightGBM and differ only with respect to the set of features included. Standard deviations are reported in parentheses.

Table B.8. Improvement in Predictive Accuracy Using Alternative Image-Feature Extraction Methods

Model	Non-Image Features	Image Features	Out of Sample R^2	Improvement over Baseline
RGB Features	Category, seasonality, price, color labels	RGB	44.04 (0.20)	+6.61%
HSV Features	Category, seasonality, price, color labels	HSV	44.30 (0.14)	+7.24%
ORB Features	Category, seasonality, price, color labels	ORB	43.53 (0.25)	+5.37%

Note: Standard deviations are reported in parentheses.

Table B.9. Online Sales and Return rates, Offline Sales and Return Rates, and Model-Predicted Online Return Rates by Product Category (based on all sales). Models estimated for categories with at least 400 items with ≥ 20 sales.

Category	Online Sales	Online Returns	Online Return Rate	Offline Sales	Offline Returns	Offline Return Rate	Number of Products ≥ 20 Sales	Predictive Accuracy Online, Out of Sample R^2 (Benchmark)	Predictive Accuracy Online, Out of Sample R^2 (Main)
Dresses	96,754	69,626	71.96%	45,923	1,615	3.52%	759	28.11 (0.65)	31.13 (0.83)
Shirts	80,586	39,379	48.87%	299,313	7,007	2.34%	1,213	14.78 (0.59)	24.08 (0.77)
Blouses	43,413	23,292	53.65%	104,778	2,667	2.55%	687	4.33 (0.96)	15.78 (1.00)
Pants	36,183	21,209	58.62%	103,353	3,264	3.16%	496	-1.10 (1.16)	-0.17 (1.25)
Knit	31,893	15,708	49.25%	137,227	3,889	2.83%	511	1.80 (1.24)	17.35 (1.13)
Jackets	21,304	12,228	57.40%	24,385	876	3.59%	302		
Blazer	13,190	7,627	57.82%	27,748	993	3.58%	166	-	-
Cardigans	11,315	4,167	36.83%	16,462	507	3.08%	69	-	-
Skirts	9,252	5,259	56.84%	26,884	746	2.77%	135	-	-
Coats	5,170	3,238	62.63%	1,299	49	3.77%	88	-	-
Bolero	4,867	3,367	69.18%	0	0	0.00%	41	-	-
Sweatshirts	3,862	2,170	56.19%	6,126	191	3.12%	56	-	-
Jumpsuits	1,902	1,303	68.51%	462	10	2.16%	27	-	-
Top	1,543	728	47.18%	0	0	0.00%	27	-	-
Leather	614	287	46.74%	809	34	4.20%	8	-	-

Table B.10. Online Sales and Return rates, Offline Sales and Return Rates, and Model-Predicted Online Return Rates by Color Labels† Models estimated for color-label categories with at least 400 items with ≥ 20 sales.

Color Category	Online Sales	Online Returns	Online Return Rate	Offline Sales	Offline Returns	Offline Return Rate	Number of Products ≥ 20 Sales	Predictive Accuracy Online, Out of Sample R^2 (Benchmark)	Predictive Accuracy Online, Out of Sample R^2 (Main)
Blue	84,947	49,054	57.75%	217,457	5,658	2.60%	1056	45.28 (0.28)	50.06 (0.47)
Grey	67,381	39,320	58.35%	92,119	2,787	3.02%	951	34.06 (0.58)	36.36 (0.54)
White	48,751	26,913	55.21%	118,060	3,154	2.67%	743	23.25 (0.60)	23.79 (0.87)
Red	42,708	25,749	60.29%	81,625	2,294	2.81%	542	45.80 (0.62)	50.90 (0.79)
Brown	41,540	23,557	56.71%	84,227	2,446	2.90%	590	19.09 (0.77)	24.72 (0.95)
Black	32,444	19,305	59.50%	67,663	2,028	3.00%	411	41.01 (0.53)	42.57 (0.83)
Green	10,550	6,581	62.38%	19,815	466	2.35%	144	-	-
Pink	3,539	2,118	59.85%	8,588	249	2.90%	53	-	-
Orange	3,054	1,865	61.07%	7,701	202	2.62%	53	-	-
Yellow	1,670	1,017	60.90%	2,356	64	2.72%	23	-	-
Violet	938	617	65.78%	2,273	66	2.90%	14	-	-
Several	318	151	47.48%	84,314	2224	2.66%	5	-	-

Figure B.11. Return Rate by Postlaunch Time of Purchase, Day of the Week and Month

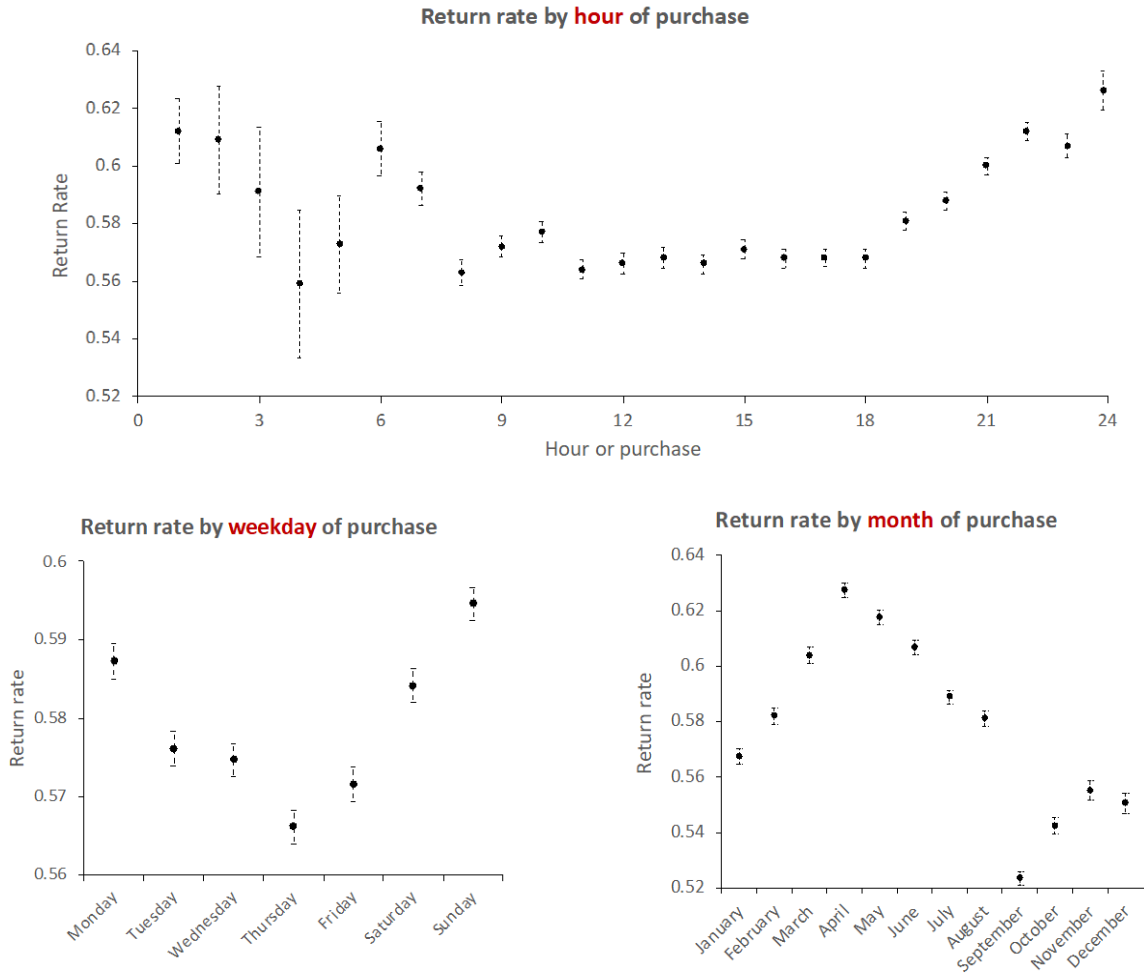


Table B.12. Product Return Rates and Price Discounts

Dependent Measure: Return rate				
	Model 1	**Model 2**	**Model 3**	**Model 4**
Proportion discounted	0.080***	0.086***	0.088***	0.078***
	(0.008)	(0.007)	(0.007)	(0.007)
Price (log10)	0.228***	0.221***	0.214***	0.279***
	(0.012)	(0.012)	(0.012)	(0.008)
Intercept	-0.156***	-0.113***	-0.083**	0.028**
	(0.044)	(0.041)	(0.041)	(0.015)
Category Controls	Yes	Yes	Yes	No
Color Controls	Yes	Yes	No	No
Seasonality Controls	Yes	No	No	No
# observations	4585	4585	4585	4585
Adjusted R-squared (model)	0.434	0.414	0.403	0.258

Note: Standard errors are heteroskedasticity robust (*$p \leq 0.1$, **$p \leq 0.05$, ***$p \leq 0.01$). R-squared is in-sample

Table B.13. Interpreting the Effect of Human-coded features (HCF) on Item Return Rates

	Regression Model	**SHAP Values**
Asymmetric	0.022***	0.92
	(0.008)	
Floral	-0.038***	-0.90
	(0.010)	
Striped	-0.063***	-0.95
	(0.009)	
Geometric/abstract	-0.020***	-0.89
	(0.007)	
Lace details	0.010	0.85
	(0.008)	
Metallic/sequin details	0.008	0.82
	(0.006)	
Graphic details	0.008	0.34
	(0.013)	
Text details	0.036*	0.69
	(0.021)	
Short Sleeves	-0.020***	-0.81
	(0.006)	
Medium Sleeves	-0.032***	-0.84
	(0.008)	
Long Sleeves	-0.032***	-0.87
	(0.007)	
Belt	0.025***	0.88
	(0.010)	
Zipper	-0.027**	-0.80
	(0.012)	
Intercept	0.039	–
	(0.026)	
Price (log10)	0.256***	–
	(0.014)	
Category Controls	Yes	Yes
Color Controls	Yes	Yes
Seasonality Controls	Yes	Yes
# observations	1,972	1,972
Adjusted R-squared (model)	0.628	–

Notes: Standard errors are heteroskedasticity robust (*$p \leq 0.1$, **$p \leq 0.05$, ***$p \leq 0.01$). Products included if sales ≥ 20.

Figure B.14. Impact of Color Clusters on Return Rates

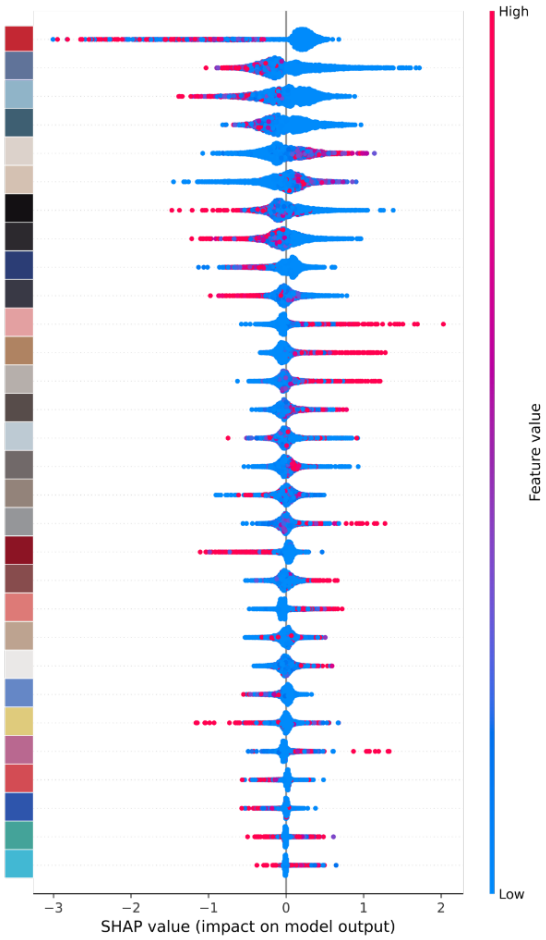
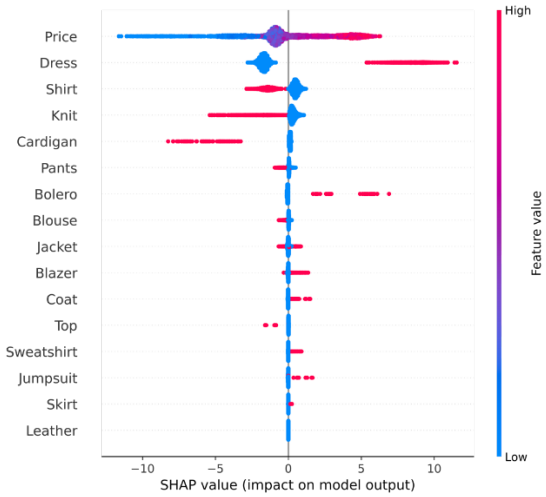


Figure B.15. The Impact of Non-Image Features

(a) Price and Category



(b) Seasonality