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

*Authors listed alphabetically

This paper has benefited from feedback obtained from the Virtual Quant Marketing Seminar Series, November 2020, Leeds School of Business Seminar Series, University of Colorado, September 2020, the Special Session on Visual Analytics, 41st INFORMS Marketing Science Conference, June 2020, the European Quant Seminar Series, April 2020, the Quantitative Marketing Conference, WU Vienna, September 2019, the Marketing Effectiveness through Customer Journeys and Multichannel Management Conference, 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. In many product categories, such as the $500 billion fashion industry, experiments in real time are not feasible because the fashion season is over before sufficient return data are observed. We demonstrate that posted fashion-item images enhance return-rate selection of assortments. We develop three interconnected models: (1) a machine-learning model to predict return rates using images and other data available prelaunch. The model predicts well; robustness tests suggest it's hard to find a better-predicting model, (2) an optimal policy to maximize profit given the imperfect predictive model, and (3) an interpretable model based on automatically-extracted image-processing features. The interpretable model provides valuable insights with which to select and design fashion items for the website. Using data from a large European retailer (over 1.2 million transactions for nearly 10,000 fashion items), we demonstrate that machine-learning methods are practical, scale to large collections and repeated fashion seasons, and improve profit relative to models using non-image data. We illustrate visually how automatically-extracted features affect return rates. Finally, we illustrate how data available postlaunch help manage return rates.

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

1. Introduction

As online retail becomes more common, many firms now sell their products through both websites (online) and bricks-and-mortar (offline) channels. Websites have many advantages over the bricks-and-mortar stores, including broader reach, lower travel costs for consumers, and saved costs of renting and operating retail space. However, a large cost for websites, relative to traditional brick-and-mortar stores, is the cost of processing product returns. 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). Even large online retailers such as Amazon struggle with managing product returns (Wall Street Journal 2018).

Product returns are vastly more common online than offline. According to our data, in which the retailer carries the same products on their website (online) as they do in their brick-and-mortar stores (offline), the average return rate per item is 53% online, but only 3% offline. The cost of processing an individual return is substantially higher for online sales. Offline, the customer brings the item back to the store and a sales associate evaluates its condition and processes the return. However, online, the firm pays to ship the item to a return processing center, and an employee opens the box, evaluates the item, and issues a refund to the customer. Additionally, the firm accrues a refurbishing fee and many returned items are discarded or sent to outlet stores. The resulting costs range between $6 and $20 per returned item (The Economist 2013). As a result, retailers' profits from online channels are highly sensitive to product returns.

In the $500 billion fashion industry, the effect of return rates on profitability is magnified. A small percent change in return rates increases net profits significantly. Furthermore, by the time a fashion retailer observes return rates, the fashion season is well underway or almost over. Fashion seasons are short and return deadlines are generous. The retailer needs to predict return rates using only data available prelaunch. The retailer also benefits from interpreting the characteristics of the items likely to be returned and items likely not to be returned. Such insight is valuable to “buyers” who make decisions on which items to carry during a fashion season. Such insight also helps in-house “designers” who design fashion items for each fashion season.

1.1. Balancing Prediction and Interpretability

Fashion retailers manage assortments. In a bricks-and-mortar store, some items are displayed to make other items attractive, such as a red sweater matched with high-margin jeans on a mannikin. To manage their collections, retailers use “buyers” who select items they believe will be popular. Traditionally, buyers use judgment to curate assortments. Returns matter less in the offline channel because returns are few and not overwhelmingly expensive.

Following common practice, the retailer in our proof-of-concept application places exactly the same items on its website as in the bricks-and-mortar stores. But the purchasing environment is different on a website. Items are displayed singly and interactions are much less important. Returns loom large and it is relatively easy to remove items from display on the website. (If a small percentage of items are removed, we do not need to worry about overall variety.) If the retailer could predict return rates perfectly, it would remove all items with a net loss (after considering returns). To remove items, the retailer needs the best feasible prelaunch predictions. When prelaunch predictions are imperfect; the retailer needs a removal strategy that takes uncertainty into account.

An interpretable model of return rates would help buyers choose the initial assortments while taking returns (mostly online) into account. For fast fashion seasons, the interpretable features should not require consumer tests, surveys, or experiments that are costly and time-consuming. While an easily-interpretable predictive model is ideal, more likely, we face a trade-off between predictive ability and interpretability. The tradeoff suggests we need two models—one model where the dial is “turned” to maximum predictive ability risking interpretability and another model in which the dial is turned toward interpretability at some loss in predictive ability.

1.2. Managing Prelaunch and Managing Postlaunch

We focus on selecting fashion items prelaunch, but sometimes initial sales rates can be observed weeks before return rates. When such data are available in a timely fashion, the difference in sales rate on the website vs. in the bricks-and-mortar store, a metric called online-offline discrepancy (OOD), enhances the predictive ability of item return rates above and beyond that based on prelaunch data. To help manage returns postlaunch we also provide insight on time of day, day of week, and month.

1.3. Topics Addressed in this Paper

We use data from a European apparel manufacturer and retailer, which include over 1.2 million transactions over two years involving nearly 10,000 unique fashion items. The data include return rates for every item sold for both the online and the offline channels. Data, which are currently used by the retailer, include information such as product category, seasonality, price, and the retailer's categorization of color. Key to our analysis, data also contain item images. The images include the same views, no models, mannikins, or backgrounds, and are reasonably constant on features such as the ratio of image pixels to total pixels. Of note, the data are per item (return rates between 0% and 100%) rather than per item and individual customer (which would be more discrete, 0 vs. 1). Using these data, we:

- demonstrate, by example, that images contain data with which to manage return rates.
- use machine learning (convolutional neural network, CNN) to summarize image data and provide incremental predictive ability (gradient-boosted regression trees, GBRT) relative to non-image data now available to the retailer. Of those tried, the CNN/GBRT gives the best predictions on our data, but the generalizable contribution is establishing that machine learning can be used to automatically analyze images rapidly and at scale and predict important economic variables.
- derive analytically an optimal decision strategy, that uses imperfect predictions to enhance profitability. We test the optimal strategy using data-based policy simulations— the retailer

can achieve 33% of the profit improvement that would be available with perfect information.

- demonstrate that a companion model, focused on interpretability, learns scalable features automatically to provide insight to the retailer, its “buyers,” and its “designers.” For example, interpretable features highlight the effect of color, pattern, and shape. The interpretable model predicts better than that based on data now used by the retailer (and many other models), but not as well as a model focused purely on prediction.
- demonstrate that investing time and money in directly measured features coded by human judges (e.g., pattern, symmetry, graphics, text, sequins) does not improve predictive ability relative to machine-learned features. (However, such features provide insight on which items are returned, as does the retailer’s color classification.)
- test the robustness of the predictive model.
  - Alternative machine-learned-feature-based models also predict well, but not as well as a CNN-based model.
  - Alternatives to GBRTs predict well, but not as well as GBRTs.
  - Although, the analytic-model-based strategy takes precision into account, measures of uniqueness and weighting by precision (sales) do not improve predictive ability.
  - The relative ranking of models and features is the same whether we apply the model across categories (e.g., dresses, shirts), or within a category.
- provide postlaunch insights to manage return rates.

2. Related Literature

2.1. Product Returns

A rich literature in marketing and operations investigates product returns. This literature focuses on (1) the effect of a return policy on retailers and consumers, (2) means to manage returns through fees, prices and deadlines, (3) managing consumers to affect return rates, and (4) how returns differ by characteristics of items such as product category and whether or not an item is a gift.

The effect of a return policy. Anderson et al. (2009) develop an individual-level model of purchase and return used to optimize the return costs for customers, Moorthy and Srinivasan (1995) suggest a return policy is a signal of item quality, and Che (1996) and Peterson and Kumar (2015) demonstrate that customers use the return option to experience the product and reduce the risk.

Managing returns through fees and prices. 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). Because lenient return policies are often mandated by law (the European Union in our data), we take the retailer's return policy as given by law in our proof-of-concept research on the ability of prelaunch images to predict and explain returns.

Managing customers. Much research (and applications) focuses on managing customers and understanding their return behavior. For example, Petersen and Kumar (2009) describe customer return behavior and how it affects future spending. Sahoo, Dellarocas, and Srinivasan (2018) study how product reviews decrease return rates by reducing consumer's uncertainty. Other studies link product returns to factors such as prices and price discounts, 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). Our focus on items complements research on customers.

Characteristics related to return rates. Experience goods are returned more often than search goods (Hong and Pavlou 2014) and products in low-risk categories such as vacuum cleaner bags and storage boxes are returned less often than products in high-risk categories such as brassieres, cosmetics, and woman's fashion. Products purchased in a new category, a new channel, or during the holiday season are returned more often, while gifts are returned less often (Peterson and Kumar 2010).

To the best of our knowledge, the literature does not focus on distinguishing which items within a category are returned more often nor on managing the selection of items prelaunch to impact profitability. Nor does the returns literature provide much information on characteristics of items such as images.

2.2. Leveraging Unstructured Data

Images are unstructured data in the sense that each image is based on millions of pixels. Many of the papers in this literature use machine learning to summarize unstructured data and predict or evaluate demand, financial performance, consumer engagement, market structure, and creative ideas (e.g., Archak et al. 2011, Chevalier and Mayzlin 2006, Lee and Bradlow 2011, Liu and Toubia 2018, Netzer et al. 2012, Onishi and Manchanda 2012, Tirunillai and Tellis 2012, Toubia and Netzer 2017). The primary focus in this literature is proof-of-concept, for example, demonstrating that customer needs can be extracted from online reviews (Timoshenko and Hauser 2019). He and McAuley (2016), Lynch et al. (2015), and McAuley et al. (2015) demonstrate by example that images are valuable to create recommendations regarding clothing styles, substitutes, and more-accurate personalized rankings. Images are also valuable to evaluate how brands are portrayed in social media, predict restaurants' survival; and predict a person's name based on an image of their face. (Liu et al. 2020, Zhang and Luo 2021, Zwebner et al. 2017).

We adopt best practices in the analysis of unstructured images to predict return rates prior to launch and to provide interpretable insights. We develop an efficient approach that uses automatic visual feature extraction. Although the methods require initial development and model training, the resulting predictive and interpretative models are fast, easy, and inexpensive to use.

2.3. Online and Offline Channels are Complements

Wang and Goldfarb (2017) find that the presence of a physical store increases customer acquisition online and Ansari et al. (2008) and Pauwels and Neslin (2015) show customer channel migration impacts demand and cannibalization. Shriver and Bollinger (2020) find channel complementarity and an overall increase in purchase volume due to retail store openings. Our retailer sells the same products offline (bricks-and-mortar) as online (website). From our perspective, both channels exist. We do not consider removing either channel. Rather we seek to manage better the online channel.

There is precedent for analyzing online and offline channels separately. For example, Dzyabura et al. (2019) provide examples where consumers evaluate the same product differently if it is sold online vs. offline. In our postlaunch analysis (§5), we demonstrate that data from both channels can be used to predict and manage returns in the online channel.

3. Prediction: Improving Online Profits by Identifying and Managing Items with High Return Rates

3.1. 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 any reason. By the retailer's policy, products 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 and offline channels during the observation period from September 1st, 2014 until August 31st, 2016 (two full consequent 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 retailer supplied 4-6 images for each item. The images were taken by the same studio, using standardized procedures, resulting in consistent image quality. In our 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. 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)[1]. An Item takes on average 55% of the image (standard deviation 13%), which mostly depends on the product category (for example, dresses are more likely to be larger than shirts). Among all the item images, 99% have the same size (2200 x 1530 pixels).

Resulting data consist of 4,585 distinct items from fifteen different apparel categories, as categorized by the retailer. Returns from items sold via the online channel range from 13–96% (56% average for this subsample, slightly above the overall average of 53%). The cost of returns from items sold online, our focus, is well above returns from bricks-and-mortar stores (data are proprietary).

3.2. Criteria to Evaluate Predictive ability

We focus in §3 on the predictive model; §4 focuses on the interpretable model. In this section, we seek to maximize the predictive ability of the model. Anticipating the analytic model (§3.8), the optimal strategy depends, in part, on the predicted out-of-sample mean-square error (MSE), thus our primary measure is the *out-of-sample*, R^2_{model}, calculated on all items (i) in our sample (K_{all})[2]:

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

where r_i is the item return rate defined as the ratio of returned ($N_{i,returned}$) to ordered ($N_{i,purchased}$) items. The return rate per item varies between 0% and 100%.

$$r_i \equiv \frac{N_{i,returned}}{N_{i,purchased}}. \quad (2)$$

$\hat{r}_i^{model}$ is the out-of-sample predicted item return rate 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

[1] The retailer did not store images for items from older fashion seasons.

[2] We report R^2 multiplied by 100.

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 D), 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 product return rates, we exclude items that were sold fewer than twenty times. In Table B.1 we demonstrate the robustness of the results to the choice of the threshold. In Table B.2 we demonstrate the model is robust whether we take $N_{i,purchased}$ as the precision metric into account or not. (Tables B.1 and B.2 are in Appendix B)

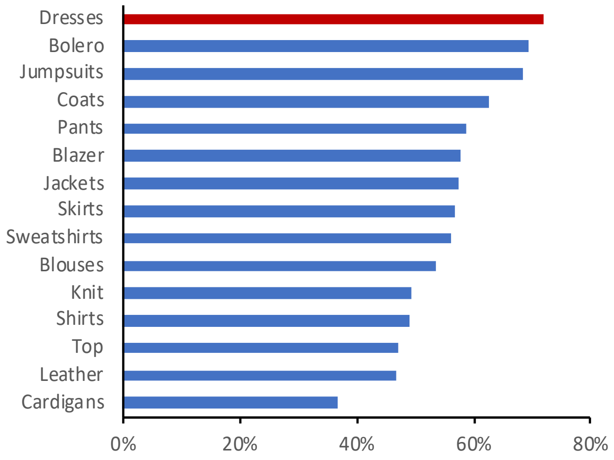
To demonstrate the robustness of our methods to different performance measures, we supplement R^2_{model} with mean absolute deviation (MAD) and U^2_{model}. U^2_{model} is a common measure used in marketing that is based on information theory (and probabilities) and measures the amount of (Shannon's) 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. Other classification metrics, such as AUC, area under the curve, require 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). Because the implications from U^2_{model} are the same as those from R^2_{model}, we simplify the tables in the text by reporting only R^2_{model}. Table B.3 reports U^2_{model}, and for completeness the mean absolute deviation of model predictions (see Appendix B).

3.3. 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**. In fashion, seasonality is clearly important and we expect return rates to vary by product category. For example, model-free evidence in Figure 1 suggests dresses (shown in red) are returned on average 72% of the time while cardigans are returned 37% of the time. See Table C.1 for sales and return rates by category (Appendix C).

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 post an item on its website. Our data do not contain sufficient information on the demand curve to optimize price. It is sufficient to demonstrate that profitability is improved when price is exogenous. Future research with improved data could include price optimization in policies to improve profitability further.

Figure 1. Online Return Rates by Category (items purchased from the retailer's website)



We can choose a variety of prediction methods with which to predict return rates. 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). We report prediction results using bagging methods (random forest) and LASSO in Table B.4 (Appendix B).

Like a regression tree, GBRT partitions the space of explanatory variables into multiple regions and predicts a value for all points in a region. The advantage of tree-based models is their ability to capture higher-order interactions among features. To avoid exploiting random variation, we regularize the tree by limiting the number of splits. A random forest generalizes regression trees by using a set of regression trees, each trained on a bootstrapped subsample of the original data. A GBRT further generalizes random forests by boosting each tree greedily based on the residual of the current model. We use LightGBM (Guolin et al. 2017) based on its performance in Kaggle machine learning competitions. This algorithm, having comparable accuracy with the alternative algorithms (for example, XGBoost by Chen and Guestrin 2016), converges significantly faster. GBRTs have performed well in marketing applications such as predicting clickstreams (Rafieian and Yoganarasimhan 2018) and predicting customer churn (Neslin et al. 2006).

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}. Our analyses are robust with respect to the variance with which the r_i are measured (Table B.2 in Appendix B).

3.4. Improving Predictions with Images (Baseline plus Color Labels)

A minimal use of images is to examine color of the product. 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. Indeed, further analysis of our data revealed that pastel colors (including pink) are returned more frequently than darker colors such as blues or blacks (Figure 6).

The retailer provides **color labels** (twelve color bins) for each item of apparel. The retailer's color labels are not perfect, for example "pink" includes many shades of pink and a single color does not fully summarize some items with multi-colored patterns or highlights. Table 1 shows that color labels improve predictions slightly relative to the baseline. While the improvement is small, the color-label model is evidence that there is information in images. We show in the next section that the GBRT/CNN model improves predictions substantially beyond predictions obtainable with the retailer's color labels. Images contain much more information than color labels for predicting online return rates.

Table 1. Baseline and Color-Label-Model Predictions

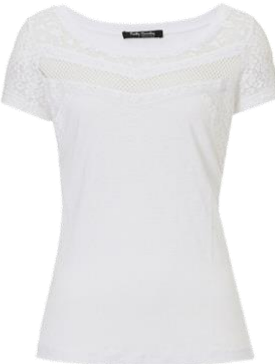
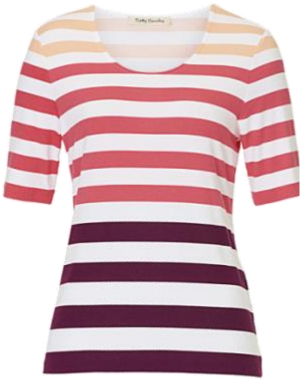
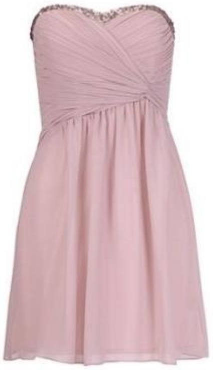
Model	**Product Features Included**	**Image Processing Features Included**	**Predictive Accuracy, Out of Sample R² (standard dev)**	**Improvement over the Non-Image Benchmark**
Non-Image Baseline	Category, seasonality, price	None	41.31 (0.18)	–
Color-labels added to baseline	Category, seasonality, price, and color labels	None	42.50 (0.20)	2.88%

Note: Models use LightGBM and differ only with respect to the set of features included. Appendix A reports U^2.

3.5 Predictions Using Deep-Learning Image-Processing 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 Benchmark 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 best. We examine the predictive ability of other image-processing features in §3.6 and more-interpretable features in §4. Details in Appendix B.

Apparel images are often more complex as illustrated by the shape of the pink dress in Table 2. Other dresses might feature floral patterns or complex 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. Each layer of the CNN transforms the features of the previous layer to obtain a new set of features. 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. For greater detail on each transformation and for an application of a CNN to unstructured marketing data, see Zhang and Luo (2018).

Our 4,500+ images are not sufficient to train a deep CNN from scratch, thus we use the second-to-last preoutput 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 this section, we seek to demonstrate the power of a deep representation of apparel images to predict returns. In §3.8 we examine whether predictive ability translates into enhanced profitability. The ResNet CNN is not the only image-processing model that does well on our data, but it is the best of those tested. To the extent we succeed with a pretuned, pretrained CNN, we provide a lower bound on what can be achieved with a custom deep-learning model if sufficient data were available.

In Table 3, we see substantial improvement with deep-learning features relative to the baseline and color-label models. We show in §3.8 that this improvement in predictive ability leads to a substantial improvement in profit when using the optimal strategy. (There is no significant improvement when we add other machine-learned features to the deep-learning features—the deep-learning features are sufficient. See §3.6) Returning to the images in Table 2, models based on deep-learning features predict return rates better for the hard-to-predict tops. Deep-learning features predict 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 prediction is within 15.2%). Before we turn to the implications of improving predictions for retailers' profitability, we examine the robustness of the model and augmentations with human-coded features.

Table 3. Predictions Using Deep Learning Image-Processing Features

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the Benchmark
Non-Image Baseline	Category, seasonality, price	None	41.31 (0.18)	–
CNN Features	Category, seasonality, price, color labels	Deep-learning	46.88 (0.19)	+13.48%

Note: Models use LightGBM and differ only with respect to the set of features included. Appendix A reports U^2.

3.6. Robustness of the CNN/GBRT Predictive Model

We must be cautious on a single data set not to try every possible set of features, models, or combinations of models. Such a combinatorial set would be quite large and potentially exploit random variation even with careful hold-out methods. Nonetheless, we would like to convince ourselves that there is no obvious candidate to improve predictive ability on our data. The Tables are in Appendix B unless otherwise noted.

Alternative predictive models. Table B.4 demonstrates that on our data the GBRT model predicts better than bagging methods or LASSO on the same features.

Within category. Table B.5 demonstrates that the models can be applied by category if the number of items in that category is sufficient. The relative ranking of the models is unchanged. Some categories are easier to predict than others consistent with the value of using category labels in the GBRT/CNN model in Table 3.

Alternative deep-learning image-processing features. Table B.6 demonstrates that on our data the ResNet CNN features predict better than the VGG-19 CNN features (Simonyan and Zisserman 2014).

Dimensionality reduction. There are 2,048 CNN features. The large number might be redundant and or strain the GBRT. Table B.7 demonstrates that linear and non-linear principal components analysis (PCA) does not improve predictions.

Alternative performance metrics. We use out-of-sample $\boldsymbol{R^2_{model}}$ as an evaluative criterion. It is possible that the relative predictive ability might change based on alternative criteria. Table B.3 demonstrates that the rank of relative predictive ability does not change if we use an information-theoretic criterion ($\boldsymbol{U^2_{model}}$) or mean absolute deviation (MAD). The analog of AUC is proportional to MSE for unbiased models and proportional to the error variance for biased models. The analog of AUC is thus redundant with $\boldsymbol{R^2_{model}}$. Nonetheless, when we compute the AUC-analog (comparisons available from the authors), the relative predictive ability does not change.

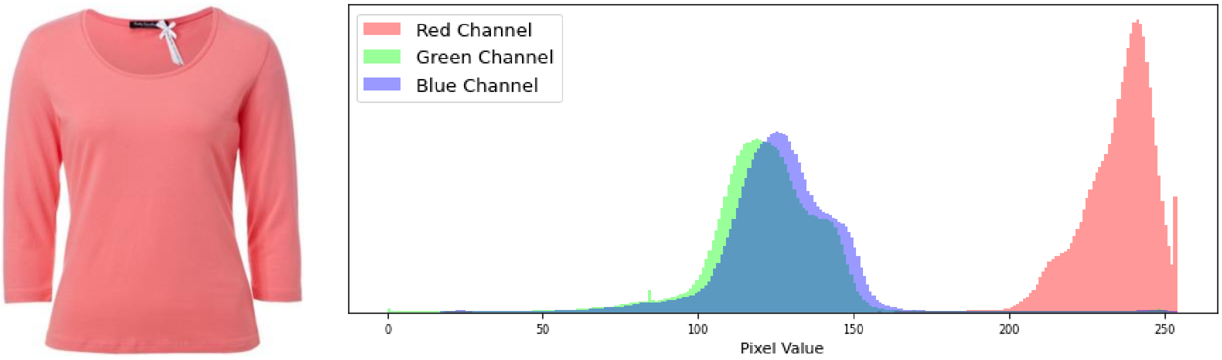
Alternative data cleaning. We screened data to require each item to have been sold in at least twenty orders. Table B.1 demonstrates that we obtain the same insights if we screen data to require a minimum threshold of either 10 or 30 times.

Precision weighting. The CNN/GBRT model predicts return rates by item. But the return rates are based on a ratio of returns to sales, thus the size of the category is not captured in $\boldsymbol{r_i}$. This might be important if return rates in larger categories are measured with higher precision. Table B.2 demonstrates that precision weighting does not improve the model.

Image of an item relative to other items. The CNN features are indexed by item, but in fashion the relative image might matter. For example, some items might be returned more if they are unique relative to the collection or if they differ substantially from previous years' collections. Table B.2 tests both uniqueness and distance from prior years, where uniqueness is Euclidean distance using the CNN features between the item and the category mean. Neither construct improves predictions.

(Black box) automated pattern & color features. Automated pattern & color features provide an alternative to deep-learning image-processing features. In fashion, a hypothesis is that the color and patterns of an item are important when predicting return rates. RGB color histograms provide one popular automated color feature. Figure 2 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 2, 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 pattern, 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). See Liu et al. (2020) for an application.

Figure 2. Example RGB Color Histogram Encoding of an Apparel Item



Tables B.5 and B.8 demonstrate that these automated pattern & color features do not predict as well as CNN-based features and that adding these features to a model based on CNN features and color labels is redundant (does not improve predictions). Table B.9 illustrates that alternative automated color features (HSV features, ORB features) do not change the basic message.

Although automated pattern & color features provide an alternative to CNN features in the predictive model, the 912 Gabor filters and 125 RGB bins do not greatly enhance interpretability. We examine more interpretable features in §4.

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. Its predictive ability does not seem to be due to chance. It is of course possible that some retailers will adopt alternative models or features for reasons outside our analysis. The performance of many alternative models is further proof of the concept of the basic idea of using deep-learning features to help manage returns.

3.7. Human-Coded Features (HCF)

Human-coded features (HCF) might have an advantage relative to deep-learning image-processing features if humans are better able to capture nuance. But the potential advantage comes at a cost. HCFs are expensive and difficult to scale for all items in every fashion season and may take sufficiently long so as to delay item selection for the fashion season. To understand tradeoffs, we collected HCFs for those features judged likely to affect returns.

Study Design. Four human judges were asked to label 2,392 images from the largest two categories of items: shirts and dresses. The 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), additional details (text, metallic/sequin, graphic, lace), sleeve length (short, medium, long, sleeveless), and the presence of belts and/or zippers. (Three judges coded sleeve length, belts, and 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 91.7%.

Results for Human-Coded Features (HCF). Table B.5 indicates that HCFs improve predictions relative to the non-image baseline (6.85%), but do not predict as well as the models with automated CNN-based features (9.92%). (The absolute predictive ability, but not the relation among models, varies when we limit the models to the two main categories.) Given the added time and cost of HCFs, the CNN features appear to be the better choice for the predictive model.

Diagnostics from HCF. While HCFs do not improve predictive ability, they are more interpretable than the CNN features. Temporarily, to make the impact of the HCFs easy to interpret, we use ordinary least squares regression (Table C.2). We present interpretations with a machine-learning metric applied to a GBRT in §4. We find that asymmetrical items are associated with a higher return rate, compared with symmetrical items. Items with pattern (floral, striped, or geometric/abstract) have a lower return rate compared with items without a pattern (solid items). Among the additional details, only text details and belts seem to be associated with higher return rates, while the presence of a zipper is associated with lower return rates. Finally, it appears that the length of sleeves is negatively correlated with the return rate. The interpretations using the metrics introduced in §4 are the same.

3.8. Using the CNN/GBRT Predictive Model to Enhance Profits Optimally.

Profits depend upon return rates. The return rate for the pink dress in Table 2 is very high. For that dress, the costs of returns likely exceed any profits earned from non-returned sales of pink dresses. On the other hand, the striped top is returned less often, and profits earned from the sales of non-returned likely exceeds the costs for the items that are returned. After we observe return rates (r_i), sales, the retailer's costs of goods, the costs of returns, and price, we can calculate whether or not it is profitable to carry the item in the website store. (Recall the retailer carries on its website only those items carried in the bricks-and-mortar stores. There are no items unique to the website.) Because the items are already inventoried for the bricks-and-mortar stores, the marginal fixed costs for carrying the items on the website are minimal. Thus, we can focus on the impact of r_i independently of the number of items sold, $N_{i,purchased}$. There are inventory and fixed costs for carrying an item, but those are well-studied, present no new insight, and can easily be added to the optimization model.

The retailer's costs for returned items are driven by a return cost (shipping and handling, c_{fix}) and by a cost that is proportional to price (returned items must be discounted or discarded because they are damaged or out of season, c_{var}). Let p_i be the price of item i and c_i be the cost to the retailer for item. If we had perfect information on return rates, r_i, we would compute profits, π_i, per item i by:

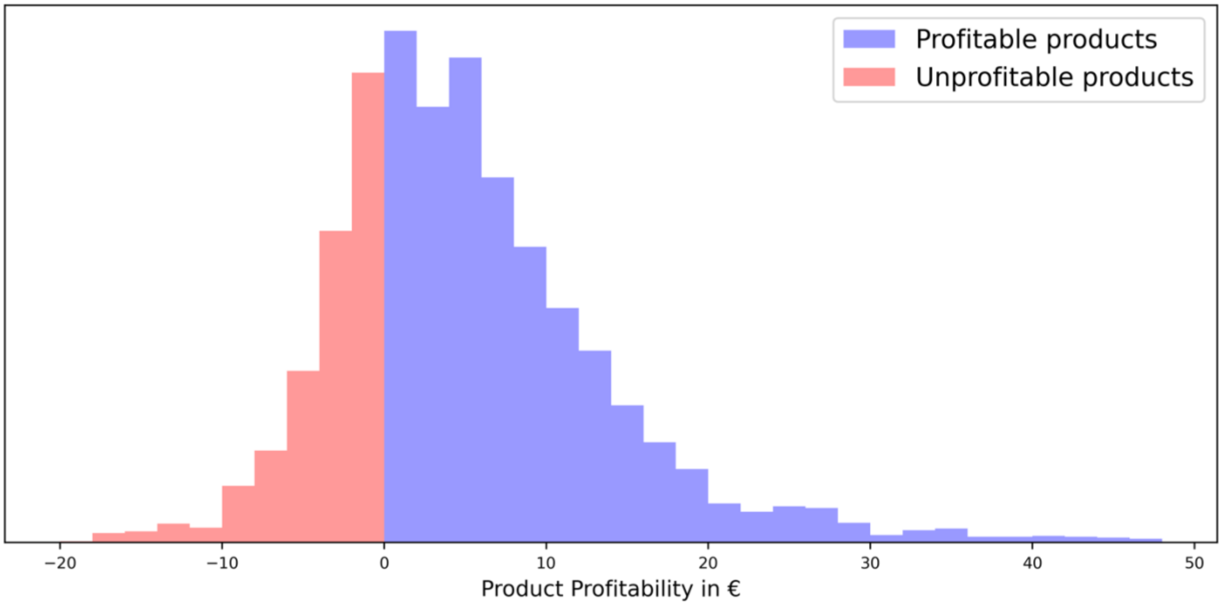
$$\pi_i = (1 - r_i)(p_i - c_i) - r_i(c_{fix} + p_i c_{var}), \tag{3}$$

Equation 3 is a linear function of the return rate, hence, by rearranging the terms, we obtain a simplified expression for the profit per item ($\mathcal{R}_i$and $\mathcal{C}_i$ are defined implicitly by Equation 4):

$$\pi_i = (p_i - c_i) - r_i\big(p_i - c_i + c_{fix} + p_i c_{var}\big) \equiv \mathcal{R}_i - r_i \cdot \mathcal{C}_i \tag{4}$$

Figure 3 illustrates the results of applying Equation 4 to our data under perfect information on returns; 27.2% of the items are unprofitable, although some are more unprofitable than others.[3] Under perfect information on return rates, the retailer would not launch items with negative profitability (the red bars in the distribution). It might then use interpretable machine-learned features to enable its designers to create new designs, convert the designs to images, and retest until only profitable items were launched. (We assume the retailer's designers are sufficiently skilled to maintain the right variety of items.)

Figure 3. Distribution of Items' Profitability in Our Data



Optimal policy. If there were imperfect information (we replace r_i with $\widehat{r_i}$), the optimal policy may differ. For example, if the estimated return rate was no better than pure noise, Equation 4 provides no insight and the retailer would launch all products. The optimal strategy depends on the ability of $\widehat{r_i}$ to predict r_i. Because decisions are made for each item, i, we temporarily drop the item subscript, i, in this section.

[3] The retailer's costs are proprietary. To illustrate the analysis, we used a fixed return cost of 5.31€ (iBusiness 2016) and a variable return cost 13.1% of an item's price (Asdecker 2015). The retailer's actual calculations are per item.

For modeling purposes, we assume that $\hat{r}$, hence our estimate of profitability, $\hat{\pi}$, has a mean equal to the true profits with a variance based on the uncertainty in r and $\hat{r}$: $\hat{\pi} \mid \pi \sim \mathcal{N}(\pi;\ \sigma_1^2)$. We assume 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 launches the item if $\mathcal{P} = 1$ and does not launch the item if $\mathcal{P} = 0$. Let $\phi \equiv (\hat{\pi}, \mu_0, \sigma_0^2, \sigma_1^2)$, then we solve the following optimization problem:

(5) $$\max_{\mathcal{P}(\phi)\in[0,1]} \mathbb{E}[\mathcal{P}(\phi) * \pi + \left(1 - \mathcal{P}(\phi)\right) * 0],$$

In Appendix A, we demonstrate that the optimal policy is a threshold policy given by Equation 6. Equation 6 yields intuitive policies as σ_o^2 and σ_1^2 approach zero (perfect information) or infinity (no information). See Appendix A.

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

Assuming that the retailer has positive priors, Figure 4 illustrates the optimal policy. (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 most items. And (3), when predictions are extremely noisy (σ_1^2 large), launch almost all items. Figure 4 suggests that we are likely to screen out more items, and be more profitable, for the better-predicting models. Because the policy depends upon σ_1^2, MSE, and hence R_{model}^2 is the appropriate criteria with which to judge the predictive model.

Perfect-prediction policy. If there were no uncertainty in predictions, the retailer would not launch the 27.2% of items that were not profitable. The impossible-to-obtain perfect-prediction policy increases profits by 25.0%. Equation 6 suggests that the policy, $\mathcal{P}(\phi)$, depends on the uncertainty in predictions (σ_1^2), hence R_{model}^2. Profits using uncertain predictions are lower than those provided by perfect information. Once developed, the deep-learning model is easy to implement; we ask whether the optimal policy based on image data provides substantial profit improvement.

Policy simulations for the deep-learning-image-processing based policy. Using our data, we simulate the model-based policies. When the GBRT/CNN model is used to determine the data-based optimal 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 no-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. The over 8% improvement relative to the baseline improvement justifies the low time and money cost of applying the GBRT/CNN model. With further development, and perhaps in other categories, more substantial profit improvement is likely. The potential for profit improvement is even greater if we consider improving items with the interpretable model.

Figure 4. Graphical Illustration of the Optimal Policy

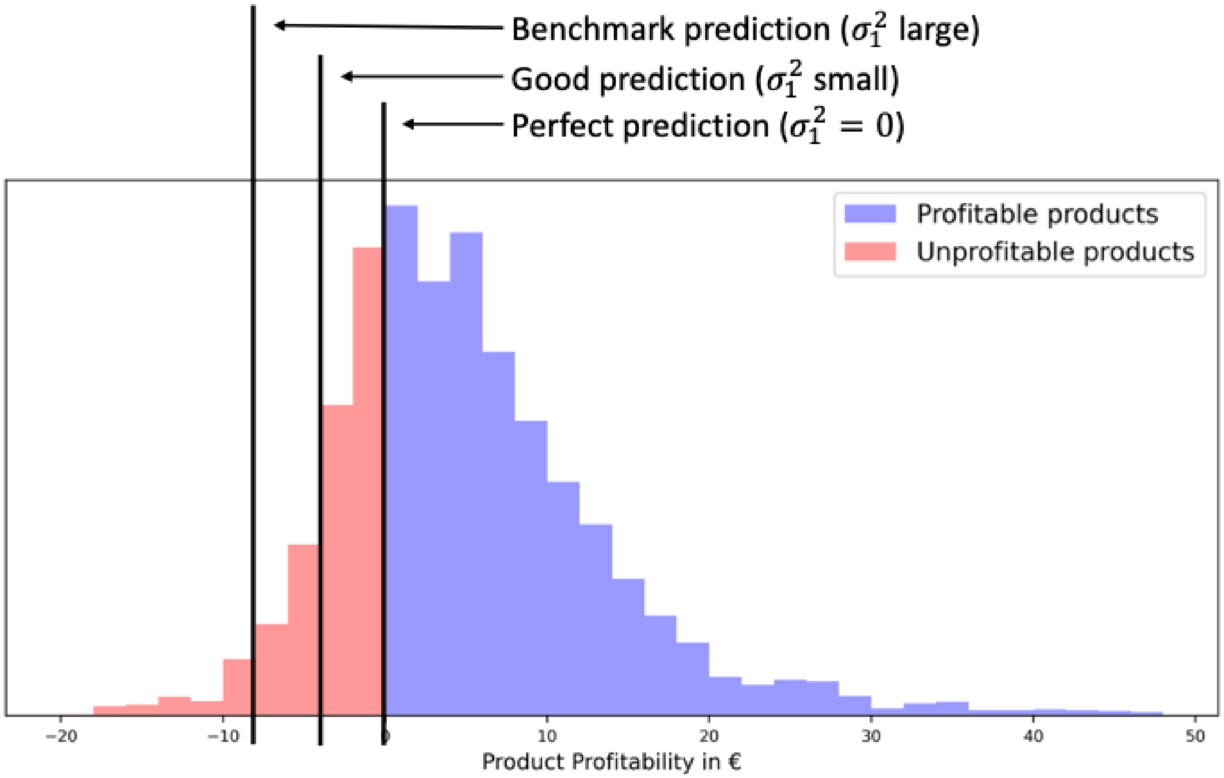


Table 4. Expected Profit Improvement Using the Deep-Learning Policy

Policy	**Estimate of Return Rate**	**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	Category, seasonality, price, and color labels	6.26% (0.13)	7.16% (0.19)
Deep-learning-based policy	Add GBRT based on deep-learning features	7.13% (0.12)	8.29% (0.23)

3.9. Summary of Predictive Model as a Means to Manage Online Assortments

Table 3 suggests that a CNN/GBRT model predicts substantially better than non-image-based model. The GBRT/CNN model once developed is easy to implement. The model is robust to alternative specification; the basic concept of using a well-specified machine-learning model predicts well, but not as well at the CNN/GBRT model (§3.6). Incurring additional costs and delays to obtain human-coded features does not improve predictions beyond that achievable with the machine-learning models (§3.7). Finally, machine-learning models, based on images available prelaunch, increases profit by 21.73% more than a non-image-based model (§3.8, 8.29% is 21.7% larger than 6.81%). Because the model, once developed, is easy to implement, the improvement in profit is valuable to the retailer.

4. Generating Interpretable Insights: Automated Extraction of Image Features

Our goal, stated in §1, is to complement the predictive model with an interpretable predictive model that helps buyers identify "red flags" to choose website assortments that avoid costly returns. Interpretable features also help designers design fashion items. In particular, we seek automatically-extracted image-based interpretable features that do not require consumer tests, surveys, or experiments. When HCFs are available, HCFs enhance interpretability.

We chose our automatically-extracted image-based features based on experience in the fashion industry. For example, pastel colors are better suited for some consumers than others, but black and darker colors tend to look good on most consumers. These rules-of-thumb suggest we might include a color feature that is more-nuanced than color labels, but still easy to interpret. Table C.3 shows this hypothesis is reasonable for our data. Similarly, we use fashion knowledge to examine other automated 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 will predict as well as the GBRT/CNN model. This is indeed the case: a model based on the interpretable features (described later in this section) 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-based baseline and the rudimentary image (color-label) model. Predictive ability is slightly better than the more-difficult-to-interpret automated-pattern-&-color-features model (§3.6) and better than the model based on HCFs (§3.7).

We made reasonable choices for metrics with which to represent the interpretable image features. Such choices are not unique, nor do they need to be unique for a proof-of-concept test. Nonetheless the high predictive ability and interpretability of the interpretable model gives us comfort. Trying many other automatic interpretable features, and their combinatoric set, would likely exploit randomness in the data.

4.1. Summarizing the Marginal Effect of the Interpretable Features

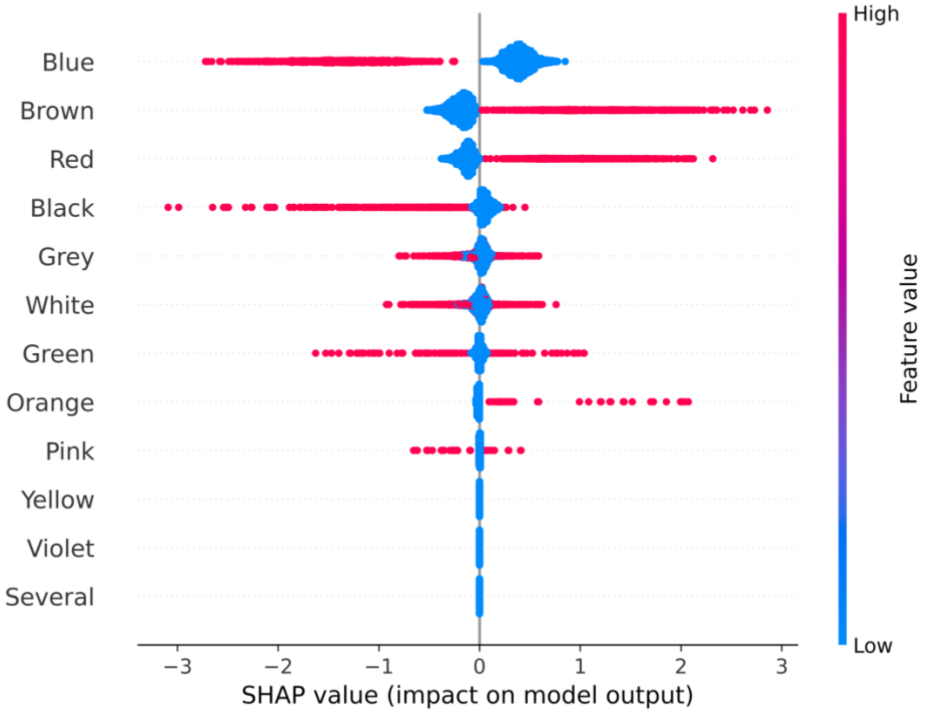
Our analyses in §3 suggest that features (image and non-image) interact and that GBRT (or another machine-learning model) is the best model to predict return rates. However, interpreting a tree-based model with hundreds of trees is challenging. Experience in the machine-learning literature suggests that the popular SHAP (SHapley Additive exPlanations) framework is a valued method to interpret the marginal impact of each feature (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. SHAP computes the marginal change in predicted return rate $\widehat{r_i}$ due to a change in the value of an interpretable image feature while taking into account all other features. By computing SHAP values for all items we obtain a sample of SHAP values. (This is often interpreted as the marginal impact given a random set of feature values.)

To determine the marginal impact of each interpretable image feature, we use the mean absolute deviation, $F_j = I^{-1} \sum_i \lfloor \phi_{ij} \rfloor$, where I is the number of items and ϕ_{ij} the SHAP values for all items i for all features j. To make the charts easy to interpret, we use Pareto charts that rank the features by F_j and display the most impactful features first. The F_j tell the retailer's buyers and designers which features impact returns the most but not the direction of impact. We complement F_j with the correlation between SHAP values and standardized feature values to indicate direction of impact.

To provide further insight, we use "bee swarm" charts to visualize the SHAP values, ϕ_{ij} for all items i for all features j. Bee swarm charts enable the buyers and designers to study how the SHAP values and the values of the feature vary over items. The bee-swarm chart details, for each item, the impact on return rate predictions of high-vs.-low values of values of the interpretable feature. Features are ranked by absolute deviation of SHAP values (F_j). If the buyers and designers so choose, they can isolate a point in the bee-swarm chart, examine what the interpretable feature means for that item, and examine the values of the other interpretable features for that item.

Before we introduce the automatically-extracted interpretable image features, Figure 5 provides an example bee-swarm chart for color labels. For example, on average if the color label is blue, then predicted return rates decrease. Predicted return rates increase if the color label is not blue. The effect varies by item. Choosing a point in the bee-swarm chart and examining the corresponding item helps buyers and designers observe the intensity of blue for that item as well as other features of the item. In §4.2, we introduce more-sophisticated color analyses based on automatically-extracted features. Of interest, browns, reds, and oranges tend to increase return rates, while blues and blacks tend to decrease return rates (Figure 5).

Figure 5. Example Bee-Swarm Chart

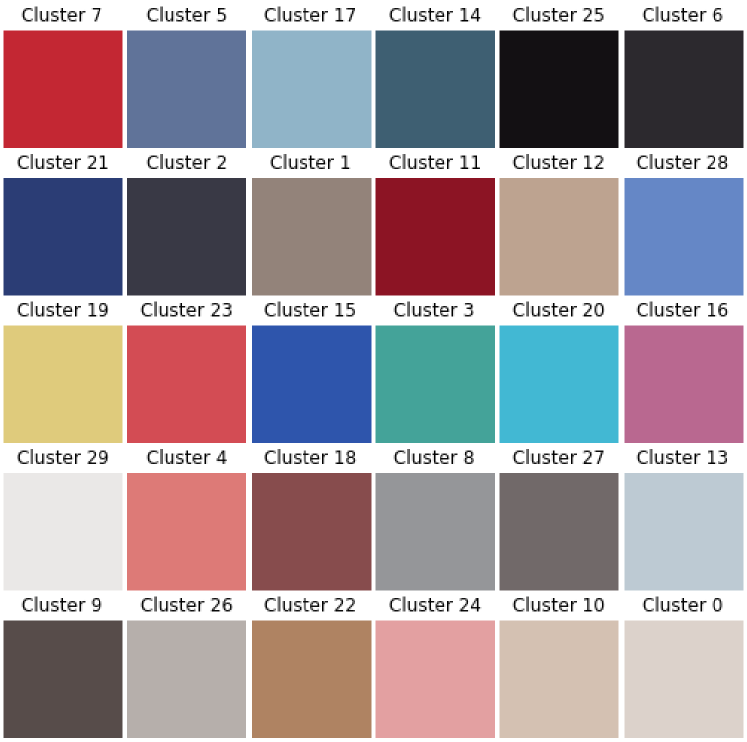


4.2. Image features

Each of the image features is based on insights, experience, and expectations from fashion apparel. 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 feature for every fashion season without costly consumer tests, surveys, or experimentation.

Color clusters. To augment the insights from color-labels (Figure 5), we use weighted K-means clusters in RGB-pixel space to visualize the basic color composition of an item (thirty clusters in our data). For each item's image, we calculate the proportion of pixels closest to the mean of the color cluster. The SHAP value for a color cluster is based on changing that proportion. Figure 6 illustrates the color cluster means ordered by average absolute SHAP value within cluster retaining the correlation sign (+/-) so that the color clusters in Figure 6 are ordered from low return rates to high return rates. Figure 6 is generally consistent the results for color labels, but the color clusters provide more-nuanced interpretations. For example, some shades of red decrease return rates, while other shades increase return rates. Blues tend to decrease return rates but the prototypical blue is in the middle.

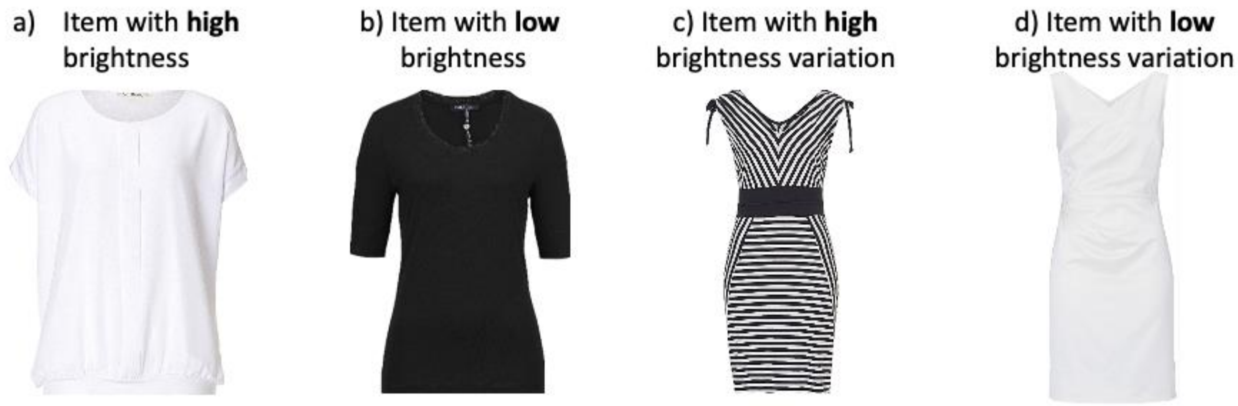
Figure 6. Color Clusters Ordered by their Impact on Return Rate (From Decrease to Increase)



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 possibly 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 a 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 7 illustrates fashion items with low and high brightness and brightness variation.

Figure 7. Illustration of Items with High and Low Brightness and Brightness Variation

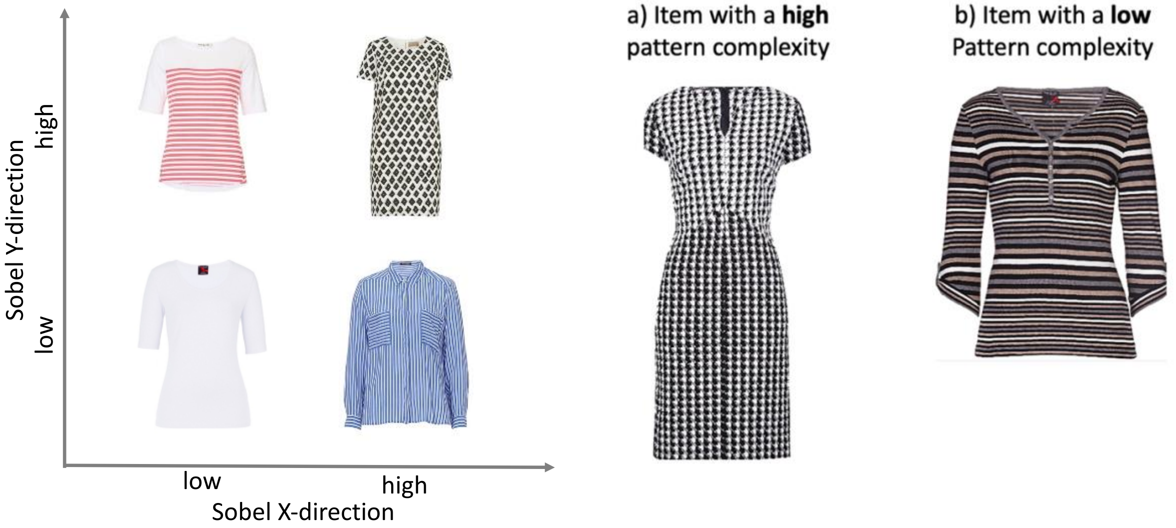

a) Item with **high** brightness
b) Item with **low** brightness
c) Item with **high** brightness variation
d) Item with **low** brightness variation

Pattern direction. Patterns matter in the automated color & pattern model, but Gabor features are difficult to interpret (§3.6). 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 left side of Figure 8.

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. The right side of Figure

8 illustrates an item (item (a)) with high complexity (lines of varying angles) and an item with low complexity (horizontal stripes with a zero angle, item (b)).

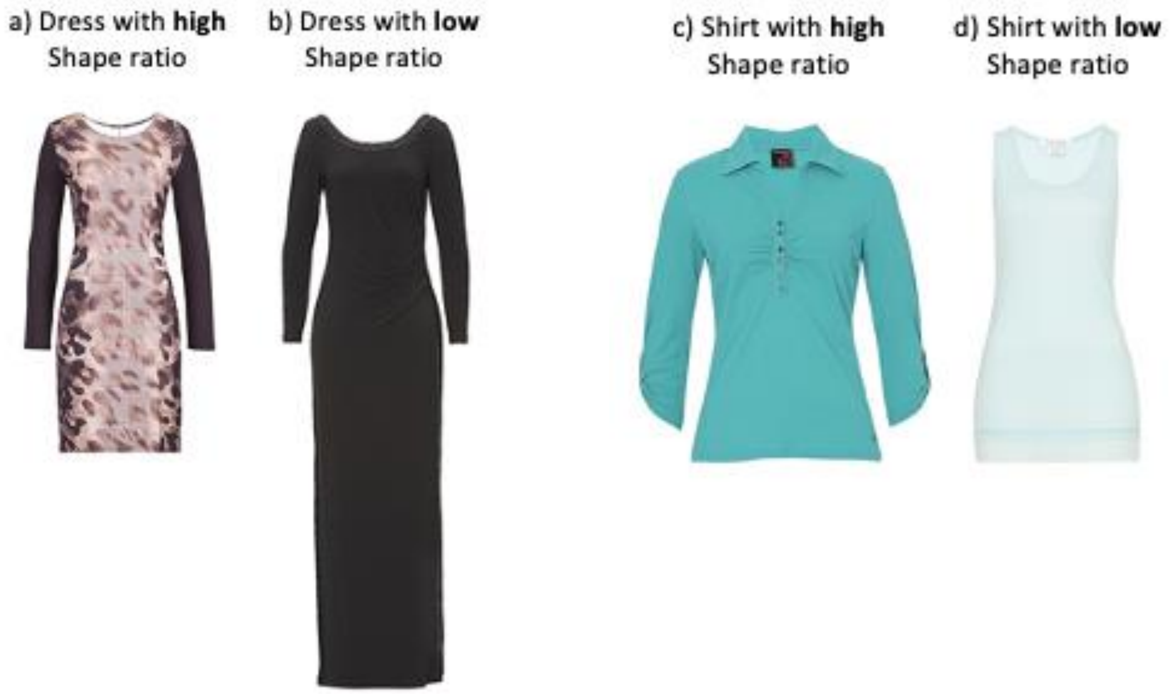
Figure 8. Illustration of Pattern Direction (Sobel X- and Y- Directions) and Pattern Complexity


Sobel Y-direction
high
low
low
high
Sobel X-direction
a) Item with a **high** pattern complexity
b) Item with a **low** Pattern complexity

Asymmetry. The human-coded-feature analysis suggests that asymmetric items have higher return rates. Shape asymmetry, horizontal color asymmetry, and vertical color asymmetry seem to matter. 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 few 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. The human-coded features, and experience in fashion, suggest that sleeve length matters. Table C.2 suggests that sleeveless items have higher return rates — all other sleeve lengths have lower return rates (Appendix C). (Consumers often evaluate how the color of sleeveless items match skin tones. This is best if the item is tried on.) The cut of an item also matters, 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). See Figure 9. **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 example figures.

Figure 9. Illustration of Shape Ratio


a) Dress with **high** Shape ratio
b) Dress with **low** Shape ratio
c) Shirt with **high** Shape ratio
d) Shirt with **low** Shape ratio

Uniqueness. Uniqueness might contain information not otherwise captured by the automatically-extracted interpretable image 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, that may have been because the CNN features already contain a (black-box) measure that captures uniqueness. For the model with interpretable image features, uniqueness does matter (§4.4).

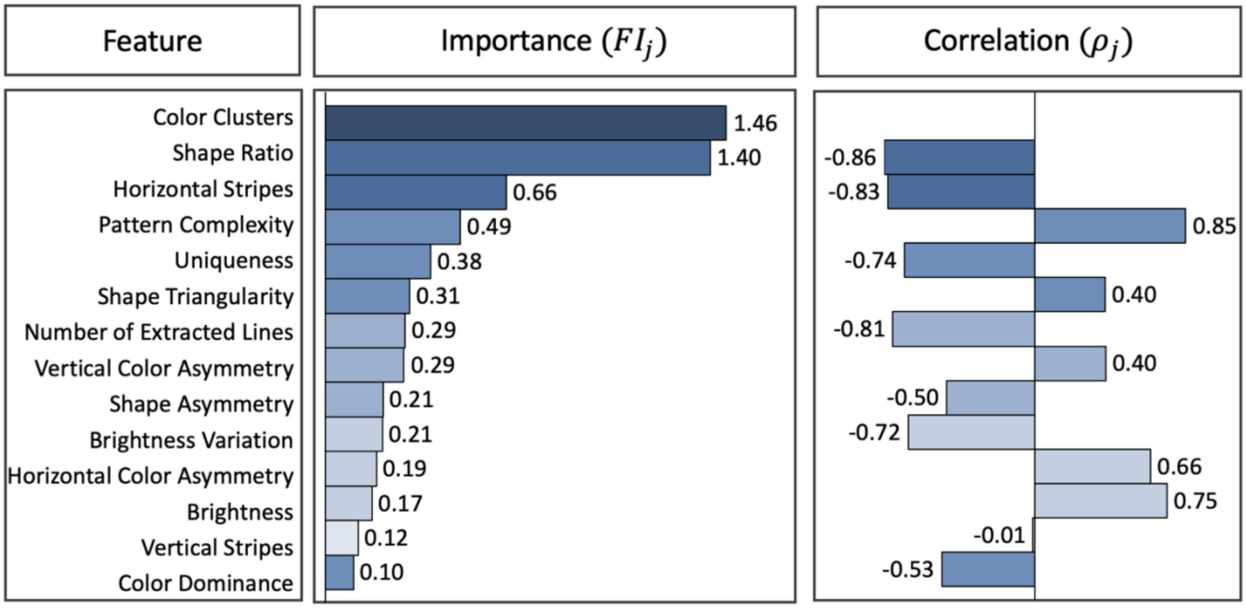
4.3. Non-Image features

We include the features from the non-image-based baseline: category, seasonality, and price. These features improved predictive ability in the GBRT/CNN model and likely improve predictive ability in the interpretation model. All three variables have been shown to relate to return rates (e.g., Anderson, Hansen, and Simester 2009, El Kihal et al. 2021). For example, in our data seasonality matters; items from the spring and summer seasons are returned at a higher rate, possibly because apparel is more visible to others in these seasons resulting in consumers having higher requirements. Figure 1 indicates that dresses have the highest average return rate and cardigans the lowest. To help interpret the impact of category, we detail online sales, online returns, offline sales, offline return rates, number of products, and the GBRT/CNN model's predictions of return rates separately for all categories in Table C.1.

4.4. Pareto Chart for Automatically-Extracted Image Features

Figure 10 provides the Pareto Chart for the automatically-extracted interpretable features. We address non-image features in the next section. Because color clusters have multiple discrete levels for each feature, we average F_j over the levels in Figure 10. We report correlations for all continuous features, but not for multilevel features such as color clusters. For multilevel features the effect on return rates varies by level, hence an average correlation would be misleading. The bee swarm charts provide more interpretable and usable insights for each level of a multi-level feature (see Appendix C for details).

Figure 10. Pareto Chart of SHAP Values for the Automatically-Extracted Image Features



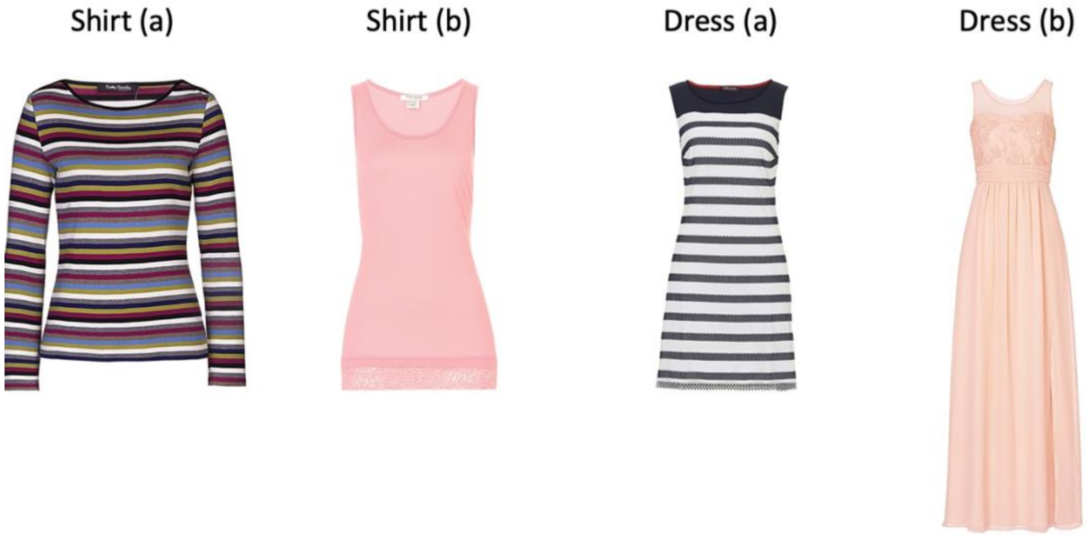
Consistent with experience in fashion apparel, color clusters have the greatest impact on return rates. Shape ratios are the next most important and the direction is as expected. More formal items (lower shape ratio) have higher return rates than more causal items (higher shape ratio). As expected, shape ratio has different interpretations for different categories (captured in the details but not Figure 10). Sleeveless dresses (lower shape ratio) are returned more often, consistent with the implications of the HCF. Interestingly, horizontal stripes are very important and decrease return rates, while vertical stripes are much less important. Complex patterns are important and increase return rates, while another important measure, uniqueness decreases return rates. Comparing Figure 10 to Table B.2 suggests that uniqueness was redundant with the CNN features in the predictive model, but not redundant in a model with interpretable features.

The brightness features are less important, likely because some brightness information is extracted by the color clusters. However, brightness increases return rates and brightness variation (many contrasting colors) reduces return rates. (This implies items with solid colors are more difficult to match to the consumer’s personal fashion fit.)

By combining the insights from Figures 5 through 10, we can predict items that are likely or not likely to be returned. For example, shirts with higher shape ratios, horizontal stripes, darker colors are less likely to be returned. Shirts with lower shape ratios, solid colors (no stripes or patterns), and a color similar to color cluster 4 are more likely to be returned. 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

(We expect and the data confirm that Shirt (a) and Dress (a) have low return rates and Shirt (b) and Dress (b) to have high return rates.)



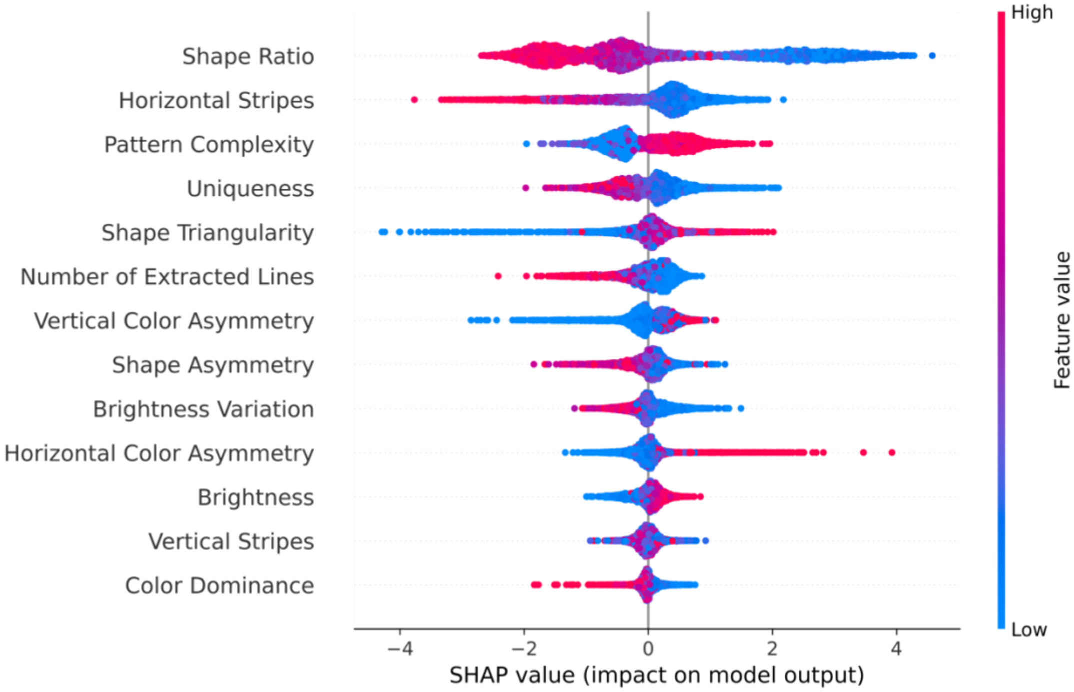
4.5. The Importance of Non-image Features

As anticipated, category, price, and seasonality are all important features with SHAP values of 2.89, 2.78, and 1.37, respectively. They are important to include in the GBRT model from which SHAP values are computed, both as controls and because of their interactions with the automatically-extracted interpretable image features (implicit in SHAP values). The non-image features also provide valuable diagnostic information as indicated in Table C.1 which provides return rates, online sales, offline sales, offline returns, predictive ability, and other data by category. Table C.7 provides bee-swarm charts for price and category and Table C.6 provides the bee-swarm chart for seasonality.

4.6. Detailed Insights from Bee-Swarm Charts

The SHAP values and correlations suggest to the retailer's designers and buyers which features matter most for managing returns. Figure 12 provides the bee-swarm chart for all image features except color clusters. (Figures 5 and Table C.4 provide the bee-swarm charts for color labels and for color clusters, respectively.) For example, on average high values of the shape ratio decrease return rates and low values increase return rates, but the effect 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.

Figure 12. Bee-Swarm Chart of Automatically-Extracted Interpretable Features for Images



Note: see Table C.4 in Appendix C for bee-swarm chart of color clusters.

4.7. Proof of Concept—Automatically-Extracted Features Provide Interpretable Insights

Figures 5 through 12 and Appendix C provide examples of how automatically-extracted interpretable image features, summarized by Pareto and bee-swarm charts, provide useful insights to retailers who seek to design and/or select items to display on their websites. Once selected, these items can be evaluated with the predictive model and decisions made with the policy model. Although the interpretable features are not unique, the features in this section provide a proof that automatic extraction is feasible based on data a retailer normally collects. Although the features required time-consuming development and testing, they are easy to implement and can be scaled to a large number of items quickly for each fashion season. Once developed, the predictive and interpretative models run quickly. Applying the CNN, the interpretable feature extraction, and both GBRT models (after training) takes less than 10 seconds per item on a MacBook Pro 2018.

5. Insights on Managing Returns Postlaunch

5.1. Time of Day, Day of Week, Month, and Price Discounts

Our focus has been on the selection of items prior to a fashion season. But our data provide insights on how to manage returns postlaunch. In particular,

- Table C.5 provides insights on how return rates vary the time of day, day of week, and month. Table C.6 provides a bee-swarm chart by month.
- Table C.8 suggests that price discounts increase returns.
  - Although we do not model endogeneity, this result suggests that prelaunch the retailer might be wary if it <u>plans</u> discounts.
  - If price discounts were known prelaunch, Table C.9 suggests the knowledge of price discounts would increase the predictive ability of the GBRT/CNN model.

5.2. Online vs. Offline Discrepancy (OOD)

By policy and regulation, item returns are observed with a time delay (weeks after purchase) and returns are a fraction of sales, albeit sometimes a large fraction. Item sales are observed more quickly and are based on a greater number of items. In most instances, reliable estimates of sales are available much more quickly than reliable estimates of return rates. When such sales data are available quickly, we might update the model with this information for better predictions and a more-profitable policy.

Online vs. Offline discrepancy (OOD) is the difference in an item's normalized sales between the website (online) and the item's sales in bricks-and-mortar stores (offline). Because total sales volumes vary between online and offline, we normalized sales to the average sales in the category to which this item belongs. We expect OOD to be correlated with online return rates because it is easier in the offline channel (vs. online channel) for the consumer to observe how the item looks on the consumer, how the item fits the consumer's fashion sense, how the item works with other fashion items, and, of course, fit. Thus, OOD might be a better predictor of website returns than returns in bricks-and-mortar stores.

In our data, return rate is positively correlated with OOD ($\rho = 0.18,\ p - value < 0.01$). Table C.1 provides online sales, offline sales, online return rates, and offline return rates by category. When we add OOD to the predictive model, $R^2_{model} = 48.66$, showing an important increase. When OOD is available quickly and reliably, it is a useful diagnostic to manage returns after launch.

6. Summary and Discussion

To the best of our knowledge, retailers do not systematically process item images to manage return rates. Our goals were to establish that retailers could predict return rates, and diagnose item return rates, based on features images that are available prelaunch. We focus on scalable automatic-extraction methods that could be used quickly and repeatedly for every fashion season and that scale to large assortments and many categories of items. The predictive model successfully predicts return rates using a criterion relevant to the optimal policy and the optimal policy improves profits. The interpretative model sacrifices a small amount of predictive ability to provide diagnostic information valuable to the retailer's buyers and designers. Both models, once developed and trained, run quickly and scale.

Like any proof-of-concept, neither the predictive model nor the interpretative model are unique. Robustness tests indicate that, of the many variations tested, it would be hard to improve predictions beyond the GBRT/CNN model. The automatically-extracted features, combined with category, price, and seasonality, explain almost as much variability as the black-box CNN features. There is little headroom for additional interpretable features, although others might do as well.

6.1. Generalizability

Our data are from the fashion industry. This industry is important by itself, but we believe that many insights generalize. Returns are important in almost all online channels, and we hypothesize that product images can be mined with deep learning and automatically-extracted image features will enhance profitability in many product categories.

6.2. Research Opportunities

There are at least five complementary research opportunities. First, the fashion retailer might improve the accuracy of the return-rate predictions and/or reduce return rates by optimizing the images to communicate key aspects of each item to consumers using methods such as those studied in Zhang et al. (2017). Second, future research might incorporate consumers' search and purchase behaviors into the management of returns—particularly, when such search and purchase behavior is item-and-category dependent. Third, image data can be investigated for postlaunch management. Image data might augment and interact with time of day, day of week, month of year, price, price discounts, and online/offline discrepancy. Fourth, we were careful to avoid overexploiting our data by not trying too many alternative models. With additional data, we might test other alternatives. Fifth, we might explore spillovers where the deletion of one item affects sales of another item.

Appendix A. Proof that the Optimal Policy is a Threshold Policy

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 $\sigma_1^2 \rightarrow \infty$ 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 $\sigma_o^2 \rightarrow 0$ 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 optimal threshold policy.: The firm solves the following optimization problem:

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

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:

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

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

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

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:

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

From (A6), it follows that:

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

Which is the threshold policy.

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

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

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:

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

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.

Appendix B. Robustness Tests for Predictive Model

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

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the 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.

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

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the 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%

Note: 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.

Table B.3. Improvement in Predictive Accuracy Using Alternative Performance Metrics

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample $U^2 * 100$ (standard dev)	Predictive Accuracy, Out of Sample, MAD $* 100$ (standard dev)
Non-Image Baseline	Category, seasonality, price	None	52.75 (0.27)	8.59 (0.01)
Color Features	Category, seasonality, price, color labels	RGB	53.72 (0.25)	8.35 (0.02)
Pattern Features	Category, seasonality, price, color labels	Gabor	53.79 (0.32)	8.36 (0.02)
Color and Patterns	Category, seasonality, price, color labels	RGB + Gabor	54.35 (0.28)	8.28 (0.01)
CNN Features	Category, seasonality, price, color labels	Deep-learning	56.70 (0.26)	8.15 (0.02)
Test of All Features	Category, seasonality, price, color labels	RGB + Gabor + Deep-learning	56.70 (0.28)	8.15 (0.02)

Note: All models use LightGBM and differ only with respect to the set of features included. AUC analog note shown because it is proportional to R^2 model.

Table B.4. Improvement in Predictive Accuracy Using Alternative Prediction Models

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the Benchmark
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%

Table B.5. Predictions for the two Largest Categories (Dresses and Shirts)

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the 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 (0.31)	+6.85%
CNN Features	Category, seasonality, price, color labels	Deep-learning	63.69 (0.19)	+9.92%

Table B.6. Improvement in Predictive Accuracy Using an Alternative CNN

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the Benchmark
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%

Table B.7. Improvement in Predictive Accuracy Using PCA (nonlinear and linear tested; linear shown)

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the Benchmark
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%

Table B.8. Predictions Using Automated Pattern & Color Image-Processing Features

Model	Product Features Included	Image Features Included	Predictive Accuracy, Out of Sample R^2 (standard dev)	Improvement over the 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%

Note: All models use LightGBM and differ only with respect to the set of features included.

Table B.9. Improvement in Predictive Accuracy Using Alternative Image-Feature Extraction Methods

Model	**Product Features Included**	**Image Features Included**	**Predictive Accuracy, Out of Sample R^2 (standard dev)**	**Improvement over the benchmark**
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%

Appendix C. Detailed Diagnostic Information for Interpretable Features

Table C.1. Online Sales and Return rates, Offline Sales and Return Rates, and Model-Predicted Online Return Rates by Product Category (based on all 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	29.82 (0.49)	33.35 (0.73)
Shirts	80,586	39,379	48.87%	299,313	7,007	2.34%	1,213	15.28 (0.47)	27.63 (0.57)
Blouses	43,413	23,292	53.65%	104,778	2,667	2.55%	687	7.19 (0.42)	15.49 (0.79)
Pants	36,183	21,209	58.62%	103,353	3,264	3.16%	496	3.23 (0.67)	5.75 (1.27)
Knit	31,893	15,708	49.25%	137,227	3,889	2.83%	511	2.18 (0.59)	16.74 (0.86)
Jackets	21,304	12,228	57.40%	24,385	876	3.59%	302	-3.94 (0.87)	-1.35 (1.78)
Blazer	13,190	7,627	57.82%	27,748	993	3.58%	166	-3.71 (1.28)	-2.10 (1.87)
Cardigans	11,315	4,167	36.83%	16,462	507	3.08%	69	1.39 (4.12)	19.80 (3.22)
Skirts	9,252	5,259	56.84%	26,884	746	2.77%	135	7.67 (1.26)	11.03 (2.30)
Coats	5,170	3,238	62.63%	1,299	49	3.77%	88	-11.74 (2.77)	0.24 (2.38)
Bolero	4,867	3,367	69.18%	0	0	0.00%	41	-9.56 (5.42)	-58.10 (8.91)
Sweatshirts	3,862	2,170	56.19%	6,126	191	3.12%	56	10.50 (1.72)	4.05 (2.48)
Jumpsuits	1,902	1,303	68.51%	462	10	2.16%	27	22.82 (3.86)	-19.05 (7.24)
Top	1,543	728	47.18%	0	0	0.00%	27	-0.79 (8.39)	11.87 (4.73)
Leather	614	287	46.74%	809	34	4.20%	8	-203.49 (9.23)	-110.29 (16.40)

Table C.2. 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
R-squared	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.

Table C.3. Online Sales and Return rates, Offline Sales and Return Rates, and Model-Predicted Online Return Rates by Color Labels†

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	46.87 (0.31)	54.50 (0.42)
Grey	67,381	39,320	58.35%	92,119	2,787	3.02%	951	36.31 (0.32)	41.24 (0.33)
White	48,751	26,913	55.21%	118,060	3,154	2.67%	743	29.77 (0.46)	32.31 (0.49)
Red	42,708	25,749	60.29%	81,625	2,294	2.81%	542	49.86 (0.26)	56.52 (0.38)
Brown	41,540	23,557	56.71%	84,227	2,446	2.90%	590	24.89 (0.36)	31.56 (0.71)
Black	32,444	19,305	59.50%	67,663	2,028	3.00%	411	42.76 (0.38)	49.45 (0.56)
Green	10,550	6,581	62.38%	19,815	466	2.35%	144	64.84 (0.40)	65.55 (0.64)
Pink	3,539	2,118	59.85%	8,588	249	2.90%	53	47.40 (0.76)	59.48 (1.40)
Orange	3,054	1,865	61.07%	7,701	202	2.62%	53	31.64 (1.32)	38.17 (2.17)
Yellow	1,670	1,017	60.90%	2,356	64	2.72%	23	31.64 (1.51)	35.65 (2.00)
Violet	938	617	65.78%	2,273	66	2.90%	14	47.24 (0.81)	52.88 (2.02)
Several	318	151	47.48%	84,314	2224	2.66%	5	19.04 (5.72)	24.75 (7.68)

† See Tables C.1 and C.2 for marginal effects in models that control for seasonality and category and/or present results by category.

Table C.4. Bee-Swarm Chart of Color Clusters

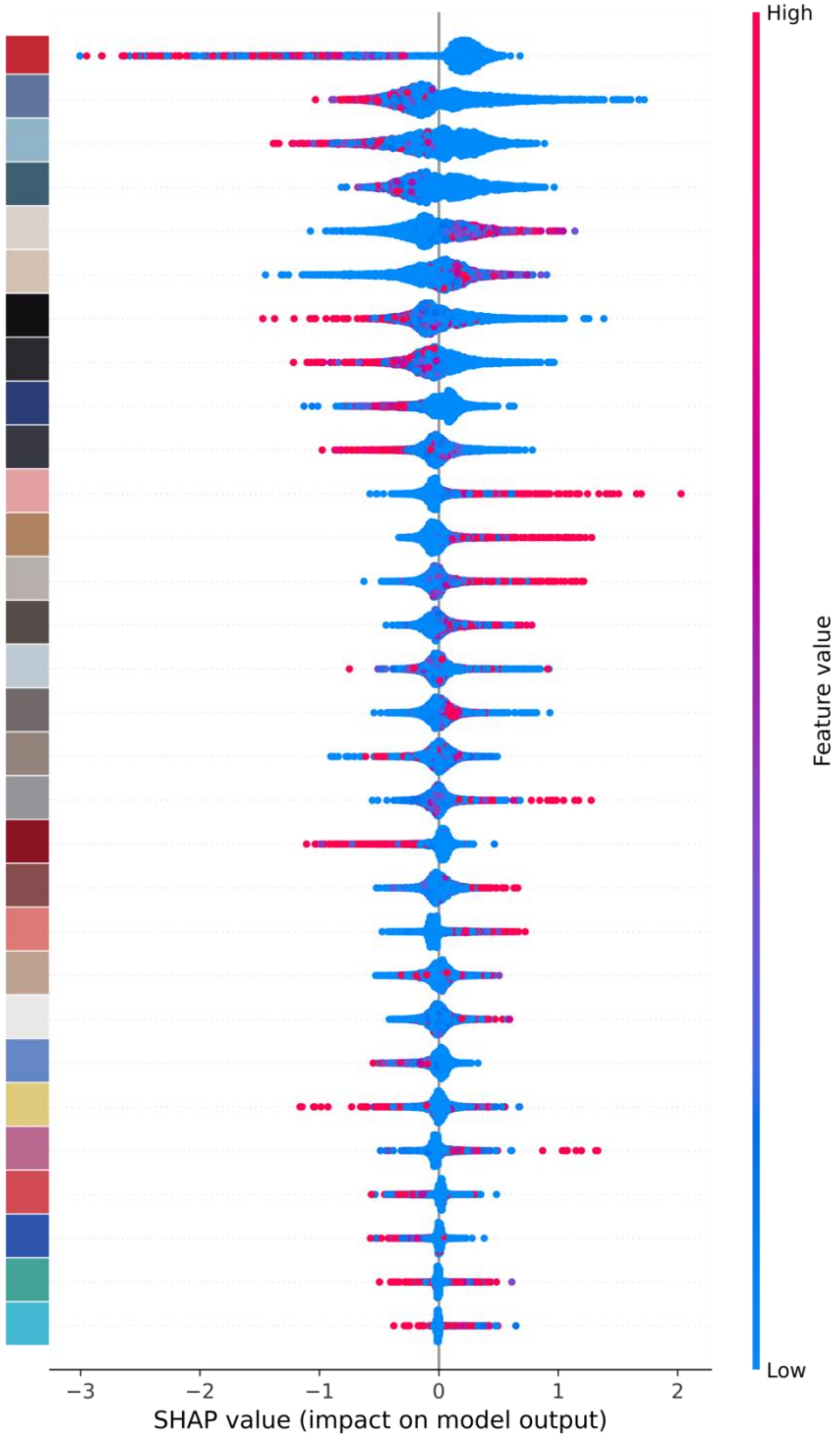


Table C.5. Return Rate by Time of Purchase, Day of the Week and Month

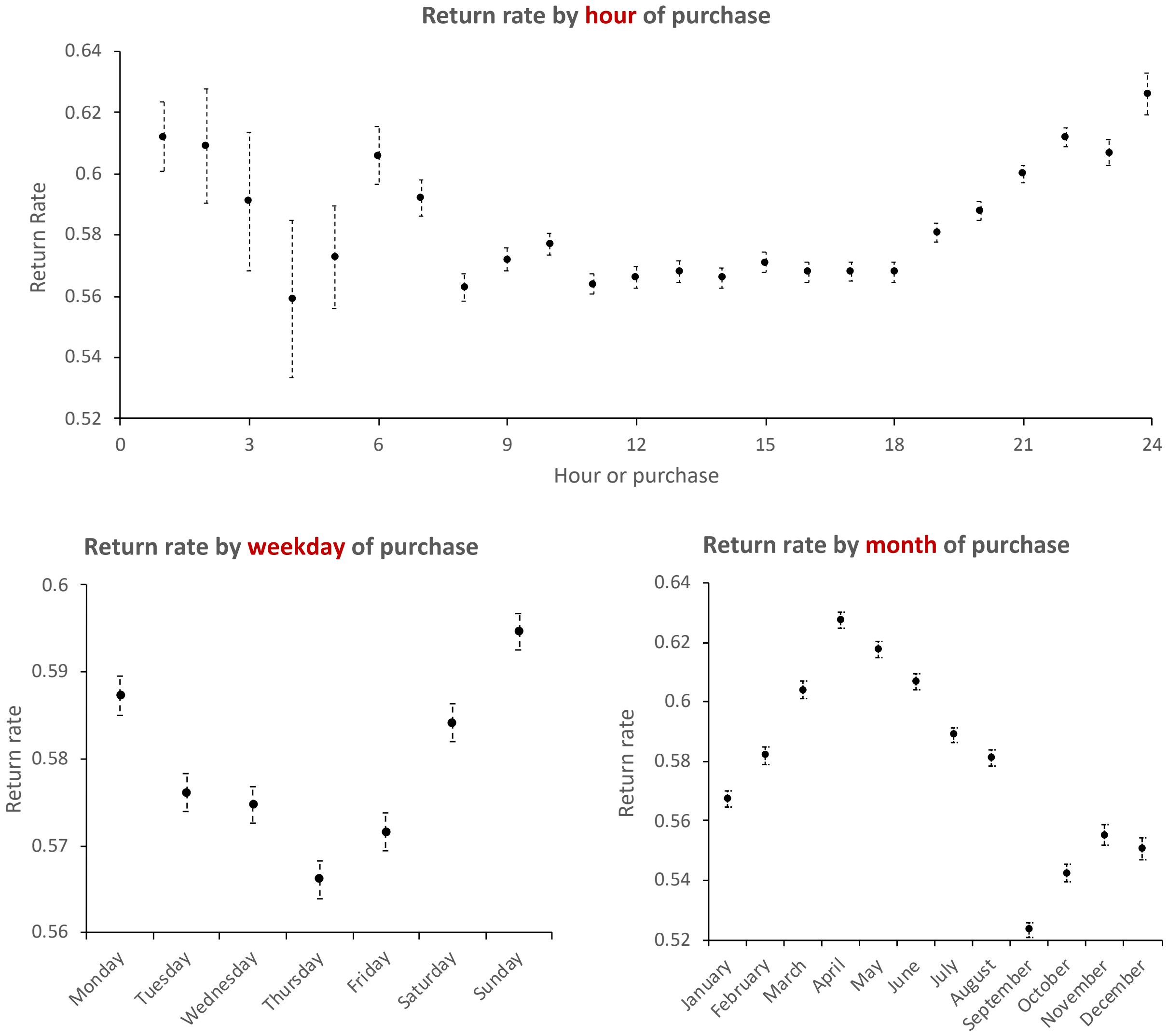


Table C.6. Bee-Swarm Chart of Seasonality (by Month)

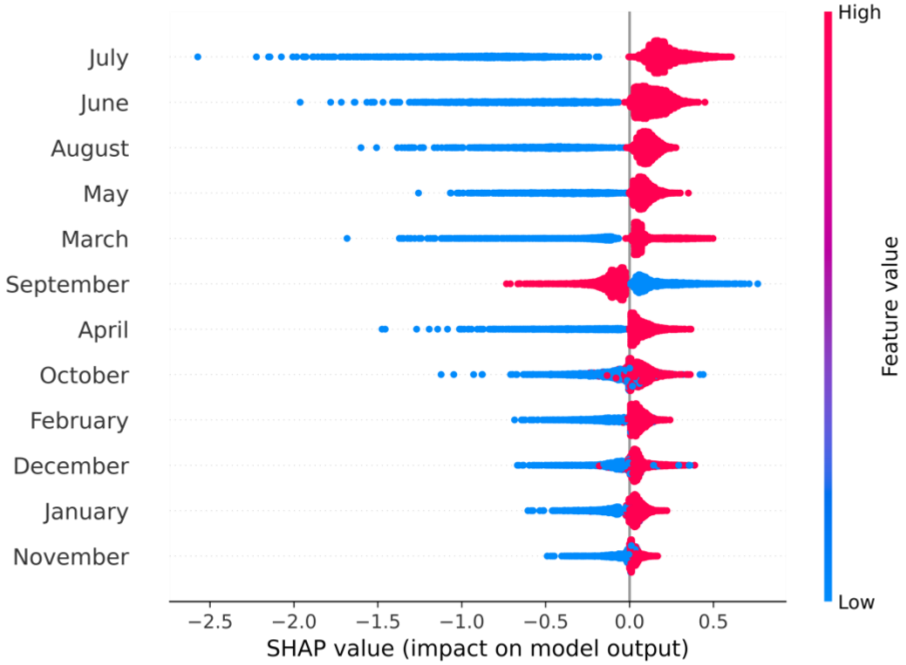


Table C.7. Bee-Swarm Chart for Price and Category

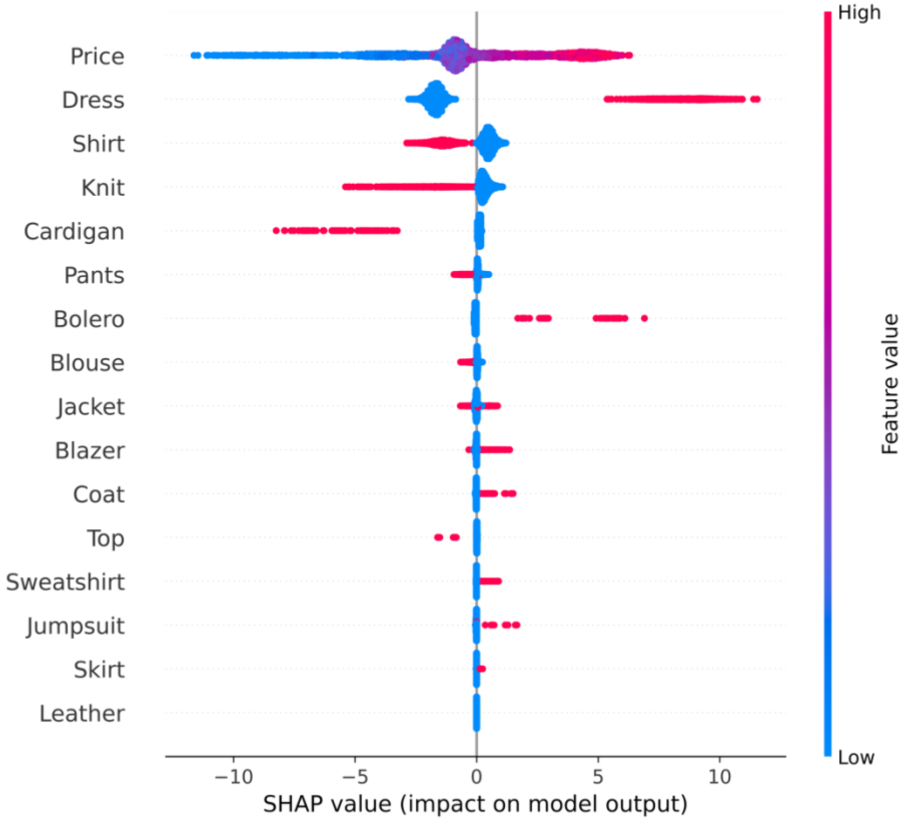


Table C.8. Product Return Rates and Price Discounts

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
R-squared	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 C.9. Predictions Using Price Discount Features

Model	**Product Features Included**	**Image Features Included**	**Predictive Accuracy, Out of Sample R^2 (standard dev)**	**Improvement over the Benchmark**
CNN Features	Category, seasonality, price, color labels	Deep-learning	46.88 (0.19)	+13.48%
CNN Features and Price Discounts	Category, seasonality, price, color labels, proportion under price discounts	Deep-learning	48.01 (0.28)	+16.22%

Note: All models use LightGBM and differ only with respect to the set of features included. Tables C.9 and C.10 treats price discounts as exogenous. Endogeneity is beyond the scope of this paper.

Appendix D. 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 D.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 D.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

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.

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

Ansari A, Mela CF, Neslin SA (2008) Customer channel migration. *Journal of Marketing Research*. *45*(1):60–76.

Archak N, Ghose A, Ipeirotis, PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management science. 57*(8):1485–1509.

Boatwright P, Nunes JC (2001), Reducing assortment: An attribute-based approach. *Journal of Marketing*. 65 (3):50–63.

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.

Briesch RA, Chintagunta PK, Fox EJ (2009) How does assortment affect grocery store choice?. *Journal of Marketing Research* 46 (2): 176–189.

Broniarczyk SM, Hoyer WD, McAlister L (1998) Consumers' perceptions of the assortment offered in a grocery category: The Impact of Item Reduction. *Journal of Marketing Research.* 35 (2):166–176.

Che YK (1996) Customer return policies for experience goods. *The Journal of Industrial Economics*, 44(1):17–24.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3):345–354.

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

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, Jagabathula S, Muller E (2019) Accounting for discrepancies between online and offline shopping behavior. *Marketing Science*. 38(1):88–106.

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*. Forthcoming.

Gittins J, Glazebrook K, Weber R (2011) Multi-arm bandit allocation indices, John Wiley & Sons, Ltd: Chichester United Kingdom.

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

Guolin K, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liw T (2017) LightGMB: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA*.

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.

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.

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.

Luo L (2011) Product line design for consumer durables: An integrated marketing and engineering approach. Journal of Marketing Research. 48(1):128–139.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. Journal of Marketing Research. 48(5):881–894.

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.

Nasr-Bechwati N, Schneier Siegal W (2005) The impact of the prechoice process on product returns. Journal of Marketing Research. 42 (3):358–67.

Narang U, Shankar V (2019) Mobile App Introduction and Online and Offline Purchases and Product Returns. *Marketing Science*. 38(5): 756-772

Neslin S, Gupta S, Kamakura W, Lu J, Mason C (2006) Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing *Research.* 43(2):204–211.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science.* 31(3):521–543.

Onishi H, Manchanda P (2012) Marketing activity, blogging and sales. *International Journal of Research in Marketing.* 29(3):221–234.

Pauwels K, Neslin S (2015) Building with bricks and mortar: The revenue impact of opening physical stores in a multichannel environment. *Journal of Retailing.* 91(2):182–197.

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.

Rafieian O, Yoganarasimhan H (2018) Targeting and privacy in mobile advertising. Working Paper.

Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: An efficient alternative to SIFT or SURF. *International Conference on Computer Vision, Barcelona*, 2011, pp. 2564–2571

Safdar K, Stevens S (2018) Banned from Amazon: The shoppers who make too many returns. *Wall Street Journal* (June 11), https://www.wsj.com/articles/banned-from-amazon-the-shoppers-who-make-too-many-returns-1526981401

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.

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

Shriver S, and Bollinger B (2020) Demand expansion and cannibalization effects from retail store entry: A structural analysis of multi-channel demand. *Available at SSRN 2600917*.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *Proceedings of International Conference on Learning Representations*.

Striapunina K (2019) Fashion ecommerce report 2019, Statista (July 2), https://www.statista.com/study/38340/ecommerce-report-fashion/

The Economist (2013) Return to Santa – E-commerce firms have a hard core of costly, impossible-to-please customers. The Economist. (December 21), https://www.economist.com/news/business/21591874-e-commerce-firms-have-hard-core-costly-impossible-please-customers-return-santa

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

Timoshenko A, Hauser JR (2019) Identifying customer needs from user-generated content. *Marketing Science*.38(1): 1–20.

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science.* 31(2):198–215.

Toubia O, Netzer O (2017) Idea generation, creativity, and prototypicality. *Marketing Science.* 36(1):1–20.

Wang K, Goldfarb A (2017) Can offline stores drive online sales? *Journal of Marketing Research. 54* (5):706–719.

Weitzman, ML (1979) Optimal search for the best alternative. *Econometrica.* 47(3):641–654.

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 (2021) Can user generated content predict restaurant survival: Deep learning of Yelp photos and reviews. Working Paper.

Zhang J, Krishna A (2007) Brand-level effects of stock keeping unit reductions. *Journal of Marketing Research.* 44 (4):545**–**559.

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.

Zwebner Y, Rosenfeld N, Sellier A, Goldenberg J, Mayo R (2017) We look like our names: The manifestation of name stereotypes in facial appearance*. Journal of Personality and Social Psychology.* 112 (4): 527–554.