Enhancing Product Categorization For Automated Tagging of Fashion and Home Products
CSE 158 Assignment 2

1. Introduction

Fashion recommendation systems have become integral to enhancing user experiences on e-commerce and social media platforms. While traditional approaches often rely on isolated product images, recent research emphasizes the importance of contextual information from scenes to improve recommendation accuracy. This study builds upon the work of Wang-Cheng et al. (2019) in "Complete the Look: Scene-based Complementary Product Recommendation," which leverages scene data from platforms like Pinterest to suggest complementary fashion items based on bounding box annotations and scene-product pairs. This approach marked a significant advancement by incorporating scene-level context into the recommendation process.

Despite such progress, a critical challenge remains: **accurately categorizing products within diverse scene contexts, where the same item can belong to different categories depending on its environment or purpose.** Addressing this issue is essential for improving the reliability of recommendation systems and enhancing user satisfaction.

Building on these insights, recent advancements such as the "Scene-aware Fashion Recommender System" (SAFRS) by Tangwei Ye et al. (2023) demonstrate the potential of integrating scene-aware items, outfits, and recommendations using transformer encoder architectures. These developments highlight the growing capability of systems to capture nuanced relationships between products and their contexts. Furthermore, work like Parmar et al.'s (2021) "Buy Me That Look" explores related problems such as product similarity, which shares methodological overlap with categorization tasks.

In this study, we aim to tackle the challenge of scene-aware product categorization using datasets derived from Pinterest's "Shop the Look" initiative. These datasets, covering both fashion and home decor domains, provide scene-product pairs with bounding box annotations and category labels. By leveraging scene-level contextual information and evaluating multiple modeling approaches, we aim to enhance product classification accuracy while providing insights into how contextual data can shape the future of recommendation systems.

2. Dataset and Methods

2.1 Dataset Description

We have two different datasets, one for fashion and one for home-related items. Our datasets consist of scene-product pairs, each accompanied by bounding box information and the corresponding category.

The fashion dataset features clothing and other apparel products, while the home dataset includes interior design and home decor items.

After conducting a series of sanity checks, we found no missing values in either dataset. However, we did discover that some products in both the fashion and home categories are repeated across various scenes. STL data statistics are outlined in Table 1.

Table 1: Dataset Statistics for Fashion and Home

Dataset	Unique Scenes	Unique Products	Unique Scene-Product Pairs
Fashion	29438	38111	72198
Home	17934	41306	93274

2.2 Exploratory Data Analysis

In terms of popularity, for categories, we observed that shoes and throw pillows have the highest number of scene-product pairs within the fashion and home datasets, respectively.

Table 2: Category Counts for Home

Category	Home Count
Furniture|Chairs	14439
Home & Garden|Decor|Throw Pillows	12941
Home & Garden|Lighting|Lighting Fixtures|Ceiling Light Fixtures	11194
Home & Garden|Decor|Rugs	9342
Hardware|Plumbing|Plumbing Fixtures|Faucets	9207
Home & Garden|Decor|Mirrors	8386
Furniture|Chairs|Table & Bar Stools	6548
Home & Garden|Decor	5617
Furniture|Sofas	5234
Home & Garden|Lighting|Lamps	5231
Home & Garden|Decor|Window Treatments|Curtains & Drapes	5135

For scene context, the home dataset has a higher mean of about 5 products per scene, while the fashion dataset has an average of around 2 products per scene, with a maximum of 3.

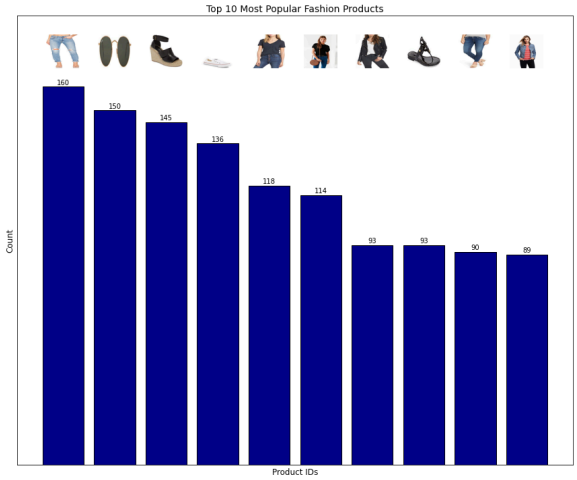


Figure 1: Most Popular Fashion Products

The following overview presents the top 10 popular items in both categories. In the fashion category, clothing products such as shirts and jeans stand out as the most favored items.

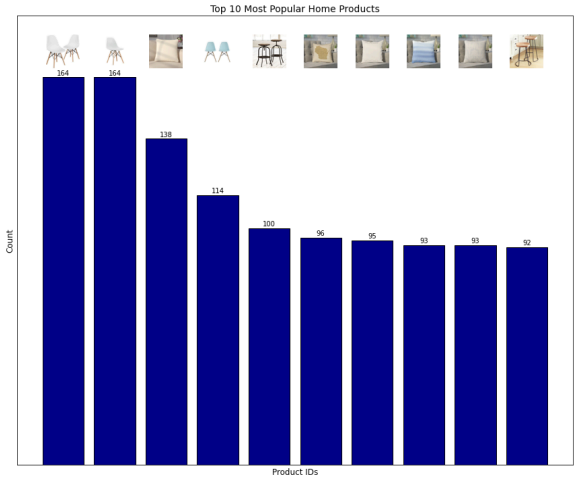


Figure 2: Most Popular Home Products

In contrast, the home category is primarily characterized by throw pillows, indicating a trend toward functional textile goods for home decor. An intriguing discovery is the presence of three chair products, all sharing an identical design style, within the top 10 home decor items. This observation prompts us to explore the factors contributing to their popularity, particularly focusing on geometric features that may influence such product categories. While pillows tend to adopt more square shapes, which are primarily picked by convention and for functionality. Some apparel items, like shirts and jeans, exhibit more unconventional forms and silhouettes which might have no impact on popularity.

3. Predictive Task

3.1 Definition

We will be focused on predicting the category of a product based on its bounding box dimensions and scene context. This task leverages the co-occurrence of products in scenes and their geometric features to infer product categories, which is crucial for automated tagging and organization of products in image-based datasets.

3.2 Evaluation Metrics

Given the nature of our dataset, characterized by multiple classes and imbalanced class distributions, we selected accuracy and F1 score as our primary evaluation metrics. Accuracy is prioritized to give additional weight to minority instances, while F1 score helps balance precision and recall amid the challenges posed by imbalanced datasets.

3.3 Baseline Models

For our baseline models from class, we chose logistic regression and Naive Bayes. We will evaluate model performance by splitting the dataset into 70% training, 15% validation, and 15% testing. We used stratified splitting to account for class imbalances, as well as used the

option ‘balanced’ for class weights in our models. The evaluation metrics from these baseline models will be compared against those from our more complex models.

3.4 Feature Engineering Approach

Following our exploratory analysis, we found a potential correlation between the dimensions of items and their respective categories. To enhance our feature set, we performed engineering on the **'bbox'** feature, calculating additional features such as the area, the center coordinates along both axes, and the diagonal length. In addition, to incorporate scene context into our models, we introduced metrics of scene diversity and density, specifically by measuring the entropy and total count of products within each scene.

3.5 Feature Selection Methods

To assess the correlation between features and the category column across both datasets, we utilized Chi-square and Random Forest methods.

The **Chi-square test** is effective for categorical target variables, quantifying the dependency between features and the target while offering a ranking based on statistical association. However, it does not account for interdependence among features. To address this, we also employed **Random Forest**, which evaluates feature contributions based on prediction accuracy, captures non-linear interactions, and is robust against noise, ensuring less significant features do not skew the results.

	Feature	Chi-Square Score
0	scene_density	5784.194939
1	relative_size	1110.064526
2	bbox_center_y	1104.918987
3	bbox_diagonal	945.498097
4	bbox_area	645.863831

	Feature	Random Forest Importance
0	bbox_center_y	0.242572
1	bbox_area	0.152075
2	bbox_diagonal	0.149025
3	bbox_center_x	0.140141
4	avg_area_per_scene	0.125064

Figure 3: Fashion | Selection Output Intersection

The output scores from the Chi-square indicate that features with high values exhibit strong dependence with category. In the fashion dataset, **'scene_density'** and **'relative_size'** exhibited the most significant relationships with the category parameter, while 'bbox_center_x' and 'avg_area_per_scene' were deemed negligible.

The Random Forest model illustrated that the most influential features for prediction included 'bbox_center_y', 'bbox_area', and 'bbox_diagonal', among others.

	Feature	Chi-Square Score
0	scene_density	76693.867955
1	bbox_center_y	3405.036862
2	scene_diversity	2831.253242
3	bbox_diagonal	2341.859352
4	bbox_area	1609.010500

	Feature	Random Forest Importance
0	bbox_center_y	0.273365
1	bbox_diagonal	0.152957
2	bbox_area	0.124856
3	avg_area_per_scene	0.120376
4	bbox_center_x	0.119785

Figure 4: Home | Selection Output Intersection

Conversely, in the home dataset, density and diversity were considered significant per the Chi-square results, yet they ranked low in importance within the Random Forest model.

The intersection of the top features can be observed in figures 3, and 4. We used this outcome of top feature sets for our baseline models.

Upon evaluating baseline models for the home dataset, the features yielding the highest validation set accuracy were 'bbox_area', 'bbox_center_y', 'bbox_diagonal', and 'avg_area_per_scene'. These features contributed to a 1% accuracy increase, achieving a total of 43.9% compared to the first baseline model of logistic regression that included all features. This suggests that while density and diversity are category-dependent, their high correlation likely skews the function, leading Random

Forest to deprioritize them as predictive indicators.

In the fashion dataset, the improvement in accuracy using all features was performed through both methodologies and their top features, ultimately resulting in an accuracy of 45.6%. The table below shows more detailed metrics.

Table 3: Fashion | Selection Output Intersection

	Accuracy		F1 Score	
Dataset	Fashion	Home	Fashion	Home
Model				
Logistic Regression	0.45	0.43	0.38	0.37
Naive Bayes	0.43	0.33	0.37	0.31

4. Model Development

We selected Random Forest, and XGBoost Classifier to begin our modeling process. We then decided to incorporate Latent Factor Features to leverage the context learning within each model. In total, we ended up with 4 final models per dataset.

4.1 Optimization process

We used **GridSearchCV** for hyperparameter tuning for all models. This step was computationally expensive, but it did leverage our results with higher accuracy for Random Forest only.

4.2 Limitations of models: scaling

The main limitation of our approach is that our dataset deals with class imbalances. As we have prevented this from impacting the model by using techniques such as stratified sampling, checking the distribution of the category count per dataset, and using a 'balanced' class weight. We cannot determine that it will reduce its class bias to zero. In the case of scalability, our models should be modified according to the nature of the user's dataset, and try other techniques such as SMOTE for minority categories if none of our approaches are effective in reducing this bias.

4.3 Strengths and weaknesses of different models

Overall, we see better performance using Random Forest rather than XGBoost. Looking into each of these models more closely, Random Forest uses a bagging approach where it constructs several trees before combining them. In contrast, XGBoost utilizes an error-correction approach, and each tree corrects previously found errors. Based on this difference, it has been shown that Random Forest may perform better for simpler datasets over XGBoost. Based on this, one weakness of XGBoost is how it is more prone to overfitting, which may be the case in our models.

Additionally, we also observed that models incorporating latent factors generally performed better over engineered features, in some of the models more than others. One of the strengths of latent factor models is that they can reduce overfitting, especially for limited samples, which aligns with our dataset's reduced scene-product count per class issue. We believe they highlight variations in scene composition, which adds into the scene-context challenge.s

5. Results

5.1 Model performance

For the **Fashion Dataset:**

- **Random Forest (RF) with Engineered Features:** Achieved an accuracy of **0.58** and F1 score of **0.55**. While the model performed reasonably well, its ability to generalize was limited by the complexity of the categories.
- **XGBoost with Engineered Features:** Showed slightly lower performance, with an accuracy of **0.56** and F1 score of **0.53**.

- ***Best model:*** **Random Forest with Latent Factor Features:** Incorporating latent factors improved accuracy to **0.61** and F1 score to **0.60**, highlighting the added value of learned scene-based context in understanding category relationships.

- **XGBoost with Latent Factor Features:** Accuracy increased to **0.60**, and the F1 score improved to **0.58**, demonstrating that latent factors positively influenced learning, but still slightly less than Random Forest.

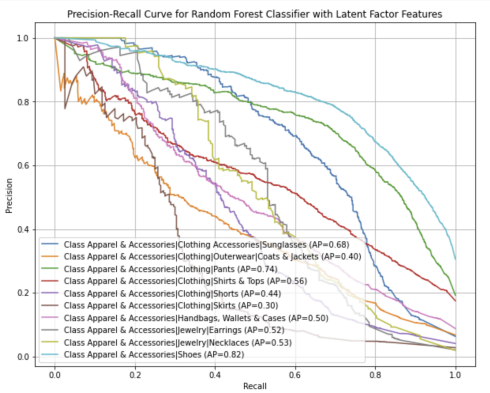


Figure 5: Best Model Precision-Recall Curve | Fashion Categories

Home Dataset:

- **Random Forest (RF) with Engineered Features:** Performed strongly with an accuracy of **0.668** and F1 score of **0.665**.
- **XGBoost with Engineered Features:** Achieved accuracy and F1 scores of **0.637** and **0.634**, respectively, slightly underperforming compared to Random Forest.
- ***Best model*** **- Random Forest with Latent Factor Features:** Showed a marked improvement, with accuracy reaching **0.73** and F1 score **0.72**, benefiting significantly from latent factor features again.
- **XGBoost with Latent Factor Features:** The model showed a modest boost in accuracy (**0.66**) and F1 score (**0.63**) compared to the base XGBoost model, though it lagged behind the performance of Random Forest.

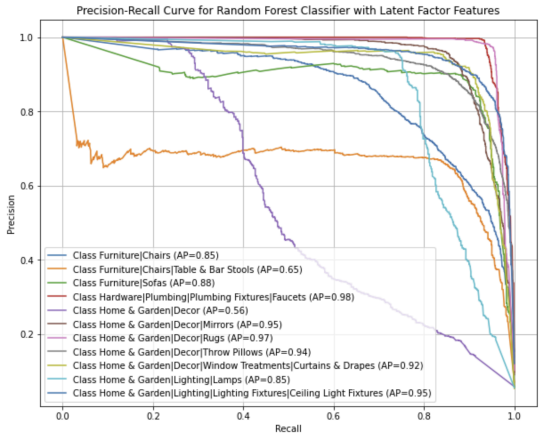


Figure 5: Best Model Precision-Recall Curve | Home Categories Validation

In Figure 5, the RF model initially showed strong performance on the home dataset, even overfitting the validation set. We addressed this issue, which resulted in lower metrics for the test set.

5.2 Discussion

Our latent factors, included scene_density, scene_diversity, bbox_area, as relatve_size,
Our top latent factors with higher coefficients for the PCA loadings were 1, and 3. Factor 1 captured the complexity of diversity features, while factor 3 the object size prominence relative to the scene.

This outcome suggests that our assumption of latent factors increasing scene-context learning in the model might be on the right path. This is why we believe that these features outperformed the other models including our baselines.

6. Literature Review

More recent research has sought to develop systems that can further improve fashion recommendations. One such novel tool, published in 2023 and entitled “Show Me The Best Outfit for A Certain Scene: A Scene-aware Fashion Recommender System”, builds upon

previous advances to create a scene-aware fashion recommender system (SAFRS), that seeks to fuse several different aspects of fashion recommendation that have been utilized previously. Namely, this paper proposes a model to incorporate scene-aware items, scene-aware outfits, and item-orient scene recommendations simultaneously, using a novel score function and a transformer encoder architecture. Using Recall@N metrics, comparison to other models, including CTL and other similar models like CSA and Outfit-T, shows that SAFRS is able to improve performance across all three recommendation aspects.

Relating to our task specifically in attempting to assign categories to products, as discussed in the paper by Wang-Cheng, et al., a similar area that this work applies to is in learning product similarities, which is a related application of categorization as essentially, both seek to compare product similarity. In “Buy Me That Look: An Approach for Recommending Similar Fashion Products” by Parmar, et al. (2021), the problem of identifying similar products is addressed with state-of-the-art computer vision modeling and object detection. Optimizing for recall@k, their model reported a R@14 of 41.7%.

7. References

Complete the Look: Scene-based complementary product recommendation
Wang-Cheng Kang, Eric Kim, Jure Leskovec, Charles Rosenberg, Julian McAuley
CVPR, 2019. https://doi.org/10.48550/arXiv.1812.01748

Tangwei Ye, Liang Hu, Qi Zhang, Zhong Yuan Lai, Usman Naseem, and Dora D. Liu. 2023. Show Me The Best Outfit for A Certain Scene: A Scene-aware Fashion Recommender System. In Proceedings of the ACM Web Conference 2023 (WWW '23). Association for Computing Machinery, New York, NY, USA, 1172–1180. https://doi.org/10.1145/3543507.3583435

Si Liu, Zheng Song, Meng Wang, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: cross-scenario clothing retrieval via parts alignment and auxiliary set. In Proceedings of the 20th ACM international conference on Multimedia (MM '12). Association for Computing Machinery, New York, NY, USA, 1335–1336. https://doi.org/10.1145/2393347.2396471

A. Ravi, S. Repakula, U. K. Dutta and M. Parmar, "Buy Me That Look: An Approach for Recommending Similar Fashion Products," 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), Tokyo, Japan, 2021, pp. 97-103, http://doi.org/10.1109/MIPR51284.2021.00022.