Data-Driven Insights and Optimization in E-Commerce

Master of Quantitative Economics

University of California, Los Angeles

Jocelyn Zulema Parra

Advisor: Randall R. Rojas

December 8th, 2023

Contents

Abstract

In e-commerce, understanding customer behavior stands out as one of the most important challenges for businesses that are aiming to improve their strategies and provide exceptional user experiences. This essay takes a look at an extremely large and complex data set, seeking to extract meaningful insights from the information we have at our disposal. With access to customer-related data that ranges from demographics and browsing habits to purchase histories, we focus on feature selection- more importantly, feature reduction. Our main question revolves around being able to determine which variables are important for effective predictive modeling. To be able to address this question, we use an approach that combines Random Forest feature importance analysis and Lasso Regularization. Furthermore, we extend our exploration to 'product type', with a particular focus on the 'Garment Lower body' category. We use Random Forests in order to determine the significance of features in this particular product category. This way, our methods help in identifying critical elements impacting consumer decisions and inclinations. Similar to Random Forest, we explore Lasso and its potential to improve generalization and reduce overfitting, providing insight into feature selection in a high-dimensional data environment. Our ultimate goal is to be able to reduce feature complexity and improve predictive modeling accuracy. We want to be able to accurately recognize if an item belongs to that of a lower garment clothing category, and from that, make inferences. The insights gained from our analysis can provide a guide for businesses, allowing them to tailor their merchandise strategies, enhance customer experiences, and make data-driven decisions. In an era defined by data abundance, this essay endeavors to sift through the digital noise, distilling the essence of customer behavior and empowering businesses to thrive in the world of e-commerce.

I. Introduction

Businesses in the e-commerce sector are constantly trying to tailor their approaches and enhance consumer experience. But to do this, they need to have a thorough understanding of client behavior. With so much information available to us in this day and age, this essay seeks to offer meaningful insights. E-commerce platforms have an extensive amount of customer-related data at their disposal, which contains demographic details like previous purchases, product preferences, and browsing patterns. The difficulty line in being able to interpret and convert this data into useful features and patterns that can actually guide business choices. With a plethora of information available, the key question becomes: where do we start, and how do we determine what is important for modeling? To tackle this question, we look into feature reduction and selection, exploring techniques such as Random Forest feature importance and Lasso classification to condense essential variables from the noise of abundant data. The main goal is to reduce the number of features in order to be able to perform predictive modeling, and in turn, those techniques allow us to condense critical features from the noise of extensive datasets, in order to obtain more accurate predictive models. Both Lasso and Random Forest are machine learning models and are regularization techniques that allow us to do classification modeling, all while preventing overfitting. Moreover, we are more specifically looking at the category within our data set titled ‘product type’, with a particular focus on the 'Garment Lower body' category. This category focuses on anything that you would wear from the waist down; that includes pants, skirts, tights, and even intimate apparel. We want to be able to accurately model whether an item in the H & M store is a ‘Garment Lower Body’. Through these approaches, our hope is to uncover the critical factors that are influencing customer choices and preferences, ultimately aiding businesses in their goals for personalized marketing and enhanced customer experiences. Not only that, but to be able to distinguish if our models are able to classify one item from another. That means, we want to be able to accurately distinguish whether a particular piece is part of a Lower Garment category based on a number of characteristics and purchasing data. Additionally, we want to draw insights from our findings. That is to say, do certain aspects influence the number of transactions taking place on a certain item, and if so, what does that mean for our analysis.

II. Data

Our dataset is a collection of customer-centric data, geared to explore the dynamics of e-commerce and customer behavior for businesses. This data was retrieved via Kaggle and was provided by H&M, for which we have access to 2 years of daily transactions. We are narrowing our data to look at the year 2019, which as a result, reduces our initial data of roughly 31 million observations, to 6 million. This is aimed at one complete year. Our dataset captures demographic details such as age. In addition, we are able to see purchasing patterns i.e., the total number of items purchased by customers and the total financial amount spent by individuals. Furthermore, we have insights on various products and categories within product types; from this, we can delve into what items customers are most interested in spending money on.

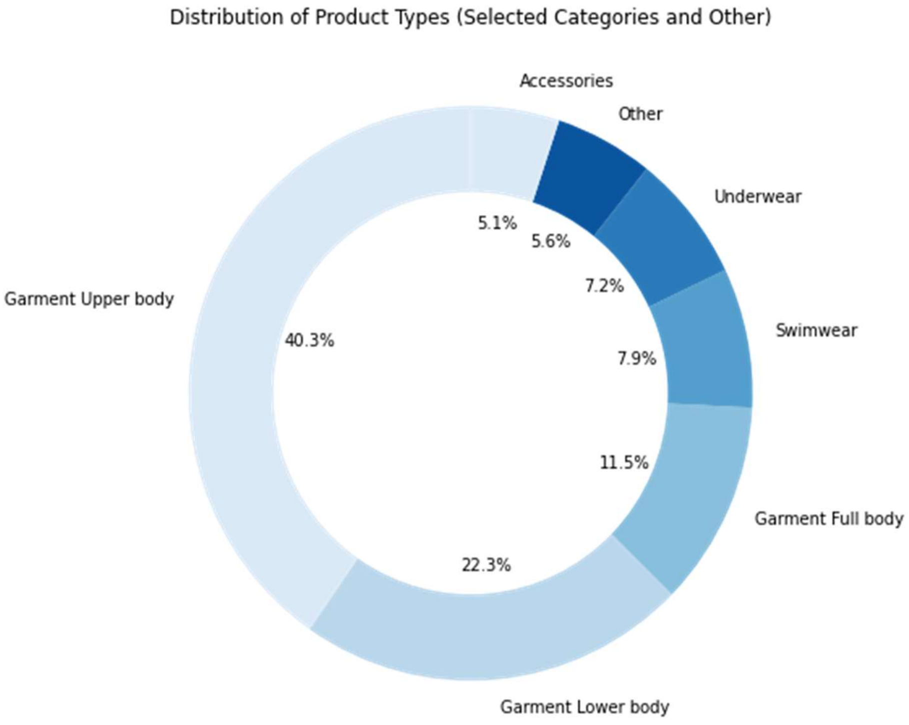

Figure 1. Distribution Of Product Types

When looking at our dataset, we can do descriptive analysis to determine what is going on. From our figure 1, we can see that when looking at the product category, we are able to visualize the top items that were purchased by percentages. That is, 22.3% of all transactions belong to lower body garments, whether those items are pants, skirts, leggings etc. However, because we have so much information, this paper primarily takes a look at garment lower body product items. Furthermore, when we delve into the garment lower body product type, we are also able to look at the top products being purchased within that grouping. In figure 2, we can see that the most bought item by far are the trousers, followed by denim trousers, followed by jersey fancy. Although the term jersey fancy and jersey basic can be misleading, the name comes from our data and they are still part of that lower garment category, but they are things like leggings or tights, shorts and even joggers. 41% of all transactions within our lower garment body are trousers. So, from this we can say that the product trouser emerges as the category with the highest revenue, standing at approximately 9.63 million, and customers purchasing these items tend to have an average age of around 38.30 years. Meanwhile, Trousers in the Denim category have a slightly lower average transaction amount of about 4.47, or about 17%, but maintain a similar customer age of roughly 38.25 years. For Jersey Fancy products (sport and athletic related lower garments), the average transaction amount is lower at approximately 2.08, and customers in this category have an average age of around 37.86 years.

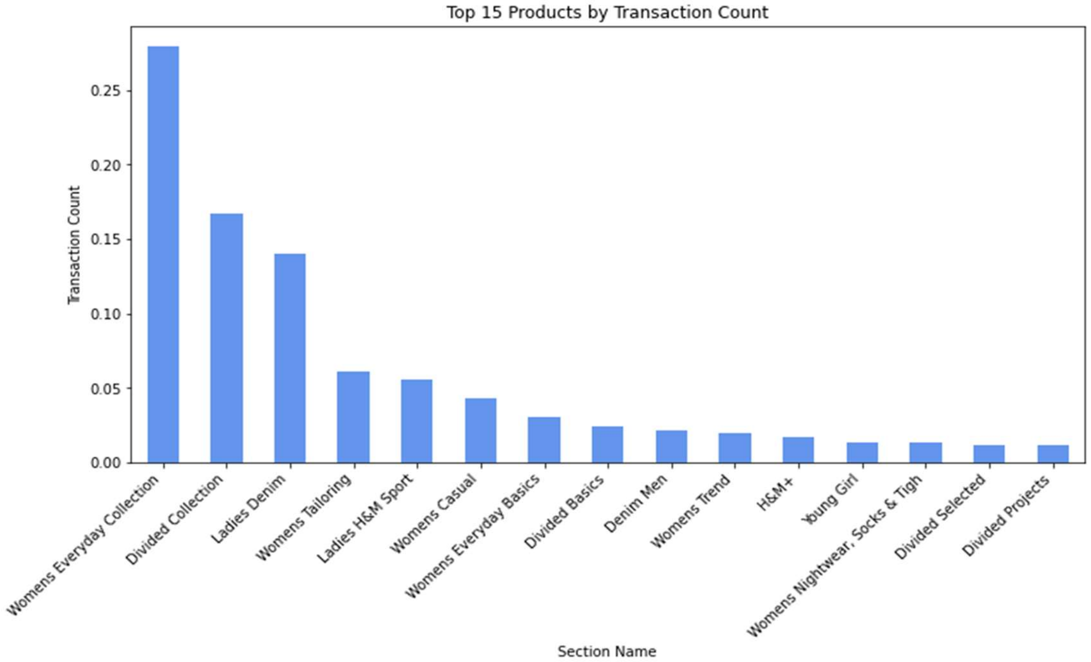

Figure 2. Top 15 Products by Transaction Count: Section

Additionally, we can look more specifically at the section name within the product category. That is, looking at what category they belong to as opposed to only looking at what kind of product that is. We have found that the highest percentage of purchases within these categories belong to the Women's Everyday Collection with around 30% of all transactions belonging to just that category.

Next in line is the Divided category with about 17% followed by Ladies Denim. We can even go as far as to say the women are buying at a higher percentage when compared to men. We can continue with our analysis and even shift our view of customer segmentation. We attempt to cluster our based on arbitrary grouping. Using a clustering algorithm, we group customers based on their purchasing patterns, revealing distinct segments that harbor unique characteristics. The clusters represent a segment of our customer base, distinguished by their age, revenue, and the number of total items bought. These clusters allow us to categorize customers into distinct groups based on shared characteristics. This segmentation is valuable for our analysis as it lets us investigate the relationships and patterns that influence customer purchasing behavior. Simply put, we aim to uncover insights into what drives customers with similar profiles to make specific buying decisions. In order to achieve this, we turn to unsupervised machine learning, specifically the K-Means clustering algorithm. Unsupervised learning is applied when we have data that lacks predefined labels or outcomes. Our goal is to assign individuals to an optimal number of clusters based only on the similarity of their attributes. To determine the ideal number of clusters, we use the 'elbow method,' which allows us to choose a value of k that provides a good trade-off between cluster quality and simplicity. Through this method, we evaluate the variance explained by different numbers of clusters. After thorough analysis, we find that the most fitting number of clusters for our customer base is 4. These four clusters represent distinct customer segments each characterized

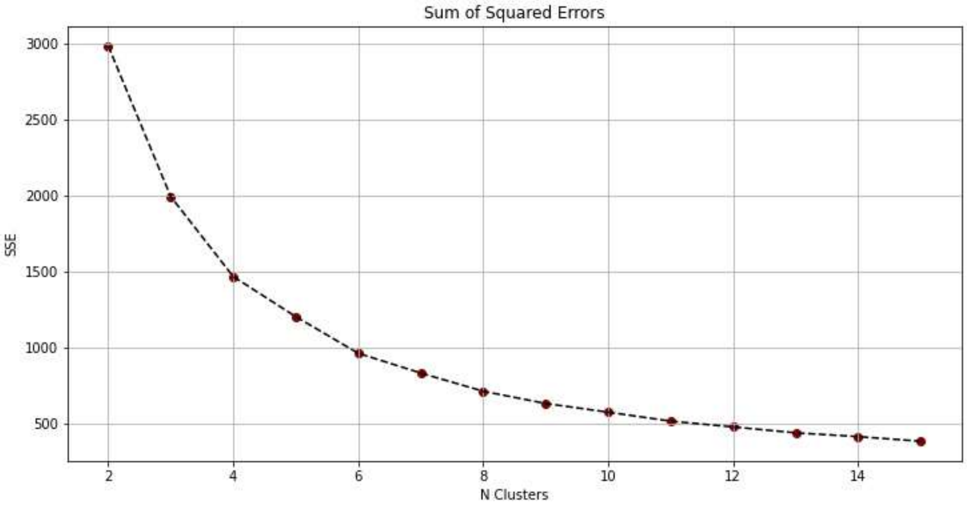

Figure 3. Elbow Method

by unique demographic and behavioral traits. We will be able to customize tactics and make data-driven decisions that address the various demands and tastes of our consumer categories thanks to this segmentation, which will be crucial in our upcoming examinations.

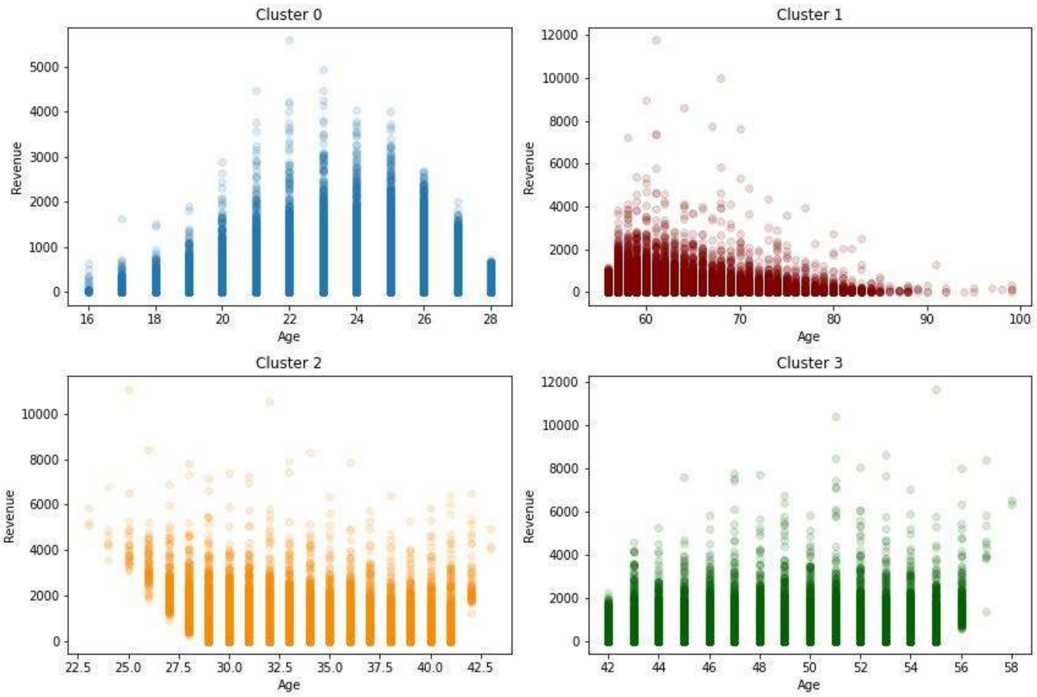

Figure 4. Cluster of Customers

From our clustering, we can see our 4 main clusters. From our figure, we can see that people in their thirties seem to spend more money compared to people in their twenties. However, when compared to people in their forties or even sixties, they tend to spend less. We can speculate that this may be due to older individuals being more established and having more disposable income when compared to younger people.

III. Modeling Approach

For the modeling aspect of our paper, we will be using a Lasso classification algorithm as well as a Random Forest. In our case, using Random Forest and Lasso over traditional linear regression can offer distinct advantages. Random Forests are great for looking into non-linear relationships and are robust to outliers, making them appropriate for datasets with complex structures like ours. They are also able to handle high-dimensional data very effectively. On the other hand, Lasso, with its L1 regularization, aids automatic feature selection. This is valuable when dealing with datasets where only a handful of features are influential or when multicollinearity is a concern. After all of that, we use 5 different metrics (Accuracy, Precision, Recall, R-Squared, and AUC) in order to evaluate our model performance. Accuracy measures the overall correctness of a model. A high accuracy score means that the model is making correct predictions for the majority of cases, and it can suggest that the model is generally reliable and performs well. On the contrary, a low accuracy score suggests that the model is making a significant number of incorrect predictions. It may imply issues with model performance and the need for improvements. Precision focuses on the correctness of positive predictions. In a binary classification, the class of interest is denoted by positive, while the opposing class is denoted by negative. The term false positives refer to wrong positive predictions, while false negatives refer to incorrect negative predictions. Similarly, true positives refer to right positive predictions. A high precision score means that the positive predictions made by the model are mostly accurate. This is important when minimizing false positive errors is crucial. A low precision score suggests that there are many false positive predictions, which can be problematic in situations where false positives have significant consequences. Recall is focused on reducing false negatives and it highlights the true positive rate. This means that a high recall rate shows us that the model is effective at identifying positive examples while a low recall score suggests that the model is missing a significant number of actual positive instances, which can be problematic in situations where detection is highly important. R-squared measures how much of the variation in y is explained by x. Here, a high R-squared signifies that the model fits the data well and explains a large proportion of the variance. It can also suggest that the model's predictions closely match the actual data. AUC, also known as Area Under the ROC Curve, measures a binary classification model's capability to differentiate between positive and negative classes across different thresholds; a low AUC score suggests that the model struggles to tell the difference between positive and negative classes, which can be problematic in classification tasks.

3.1 Lasso

We are using a Lasso approach, which aims to help prevent overfitting and improve the generalization of the model. We are implementing a data sampling approach where we randomly sample a total of 50,000 observations over 100 iterations. Our objective is to create a balanced sample. In order to do that, we prepare the dataset by creating a sample with a 50/50 (50% positive

50% neg we want a balanced sample for our analysis) split for a specified column, in this case the product group of the items, and more specifically, the category 'Garment Lower body'. These simulations highlight sample balance, ensuring an equal representation of 'Garment Lower body' product category samples (positive class) and samples from other product categories (negative class) within each iteration. This is crucial in order to gain meaningful model training. Additionally, the process introduces randomness as it randomly selects samples during each simulation, injecting variability into our model parameters. This approach helps us in assessing feature importance while considering different random samples, contributing to a more comprehensive understanding of our data.

3.1.2 Feature Selection

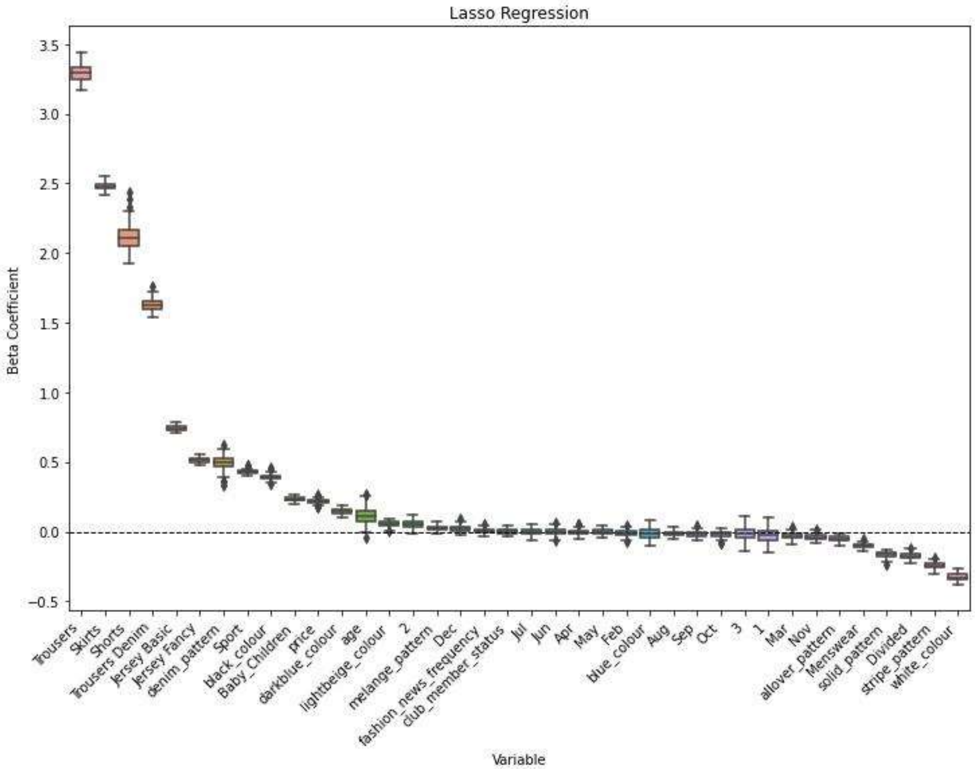

Figure 5. Feature Importance for Lasso

From our output we can see that the important features seem to be trousers, denim trousers, shorts, and skirts. After the feature "sport", we can see a considerable drop in the beta coefficients, the features near or at the 0 threshold tend to be less significant. However, if those attributes pass the 0 bound, they can be as equally significant as the variables with high betas. All of the months seem to have little to no significance in determining whether someone is buying items from the lower garment products. Based on our findings, we can speculate that it does not matter what month of the year it is, people will still buy lower garment items. We can also say that clustering may not be as significant as we thought. That is to say there are no different characteristics in our customer base, maybe we do not actually need to cater to a certain group of individuals/ customers. We see that age is near the 0-threshold signifying unimportance. This observation is expected as there might be multicollinearity because they were designed with the age characteristic despite it being one of the most important features in the clustering aspect.

3.1.3 Testing

We are using all of the total samples and 33% of all the total samples in a given iteration are reserved for the testing sample. From there, we are scaling our data, using a standard scaler from scikit learn. While doing this, we are scaling both the training and the testing sample, based on the training set parameters: the mean and the standard deviations. From our feature selection process, we determined that the important features were: trousers, skirts, shorts, trousers denim, jersey basic, jersey fancy, denim pattern, sport, black color, baby children, and white color. We then are fitting a classification Lasso since it is a way to achieve classification in a single model. Our goal is to predict what our x's will be, we are using a testing sample of x data to make predictions. In order to evaluate our models, we are using the metrics test accuracy, R^2, recall, precision, and AUC.

Table 1. Metric Evaluation for Lasso

Lasso	Average	Standard Deviation	Minimum	25th Quantile	Median	75th Quantile	Max
Accuracy	0.907558	0.002199	0.902000	0.906152	0.907455	0.909091	0.912364
Precision	0.960727	0.011175	0.923362	0.961798	0.963621	0.965621	0.968820
Recall	0.849912	0.010252	0.836639	0.844948	0.847204	0.850734	0.884503
R^2	0.630217	0.008794	0.607995	0.624583	0.629817	0.636356	0.649451
AUC	0.963094	0.001334	0.959179	0.962206	0.963044	0.963964	0.966022

We can see from our table that the lasso model has an average accuracy of approximately 90.76%, with a standard deviation of around 0.22%. The minimum accuracy in our table is 90.20%, while the maximum accuracy is 91.24%. The standard deviation of 0.22% suggests that the model's scores across different simulations only vary slightly, but those variations are relatively small-. The majority of accuracy values falling between the 25th and 75th percentiles suggests that the model constantly performs well across various simulations, with most accuracy scores falling within this range. The minimum accuracy from our results at 90.20% means that even in the worst-case scenario (the lowest level of accuracy among the simulations), the model's accuracy remains above 90%, which is generally considered good performance. To sum up, Lasso exhibits a strong and consistent ability to accurately classify data points.

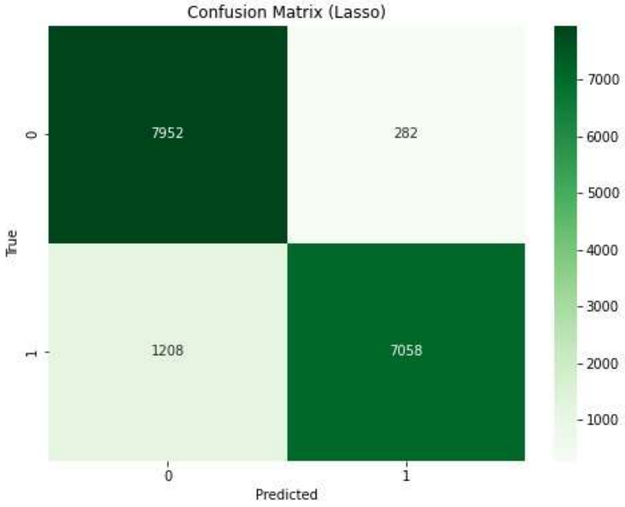

Figure 6. Confusion Matrix for Lasso

Similarly, we constructed a confusion matrix of our results to showcase the true positives and the true negatives of our classification models. Our sample size is 50,000 and we reserved 33 percent of the values for testing, which comes out to 16,500. While our metrics tell us how well we did overall, our confusion matrix gives us a more detailed view of where we went wrong. In our analysis, we correctly classified 7258 true positives and 7672 true negatives. Likewise, we have 607 false positives and 963 false negatives. Additionally, our confusion matrix gives us the necessary data to calculate the F-1 score, which comes out to be 88.2 percent. This indicated that our model is performing well in terms of both precision and recall. Overall, we are able to accurately classify whether it was a lower body garment item.

3.2 Random Forest

As stated earlier in our Lasso discussion, our dataset contains intricate and non-linear relationships between customer features and behavior. Random Forest is great at capturing these complexities that allow us to see important relationships. In addition, with a large number of features in our dataset, Random Forest can efficiently navigate high-dimensional data. This prevents overfitting and ensures our models generalize well. But most importantly, we use Random Forest because it offers a built-in feature importance measure. This feature selection process helps us pinpoint the most influential variables, allowing us to reach our goals. So, we implement a random forest for feature importance analysis. After the simulations were conducted to ensure robustness, a balanced sample of data was created that consisted of an equal number of instances from the 'Garment Lower body' category and other categories. The reason we did this was to prevent the model from being biased toward the majority class. It is important to do this, otherwise, it means that if the class we are attempting to predict is the predominant one, there's a high likelihood that our predictions will be close to or reach a value of 1 because it is the majority class. It makes sense we get near perfect predictions if our data is not balanced. Then, the features in the sampled data were standardized using StandardScaler. This step is done to ensure that all variables have the same scale, preventing any particular feature from dominating the model due to its magnitude. With that, the approach gives us a very robust assessment of feature importance across various simulations, all while keeping our data vastly balanced.

3.2.1 Feature Importance

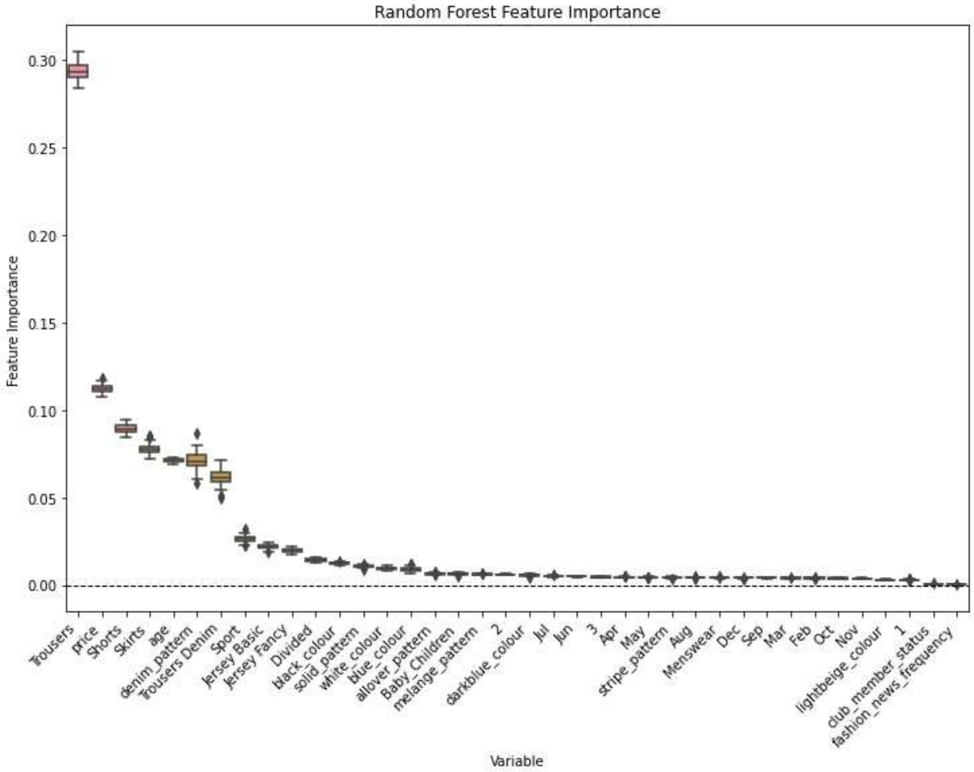

Figure 7. Feature Importance for Random Forest

From the output, we can see that the features with the most significance seem to be trousers relative to the other predictors. After denim trousers, we see the feature importance scores significantly drop. We went ahead and included months to see if there was anything significant about the time by which a customer made a purchase. However, as we can see, they are closer to 0 in the importance score, signifying their unimportance. Thus, we can conclude that our findings have significant implications for our analysis; we can now streamline our predictive modeling efforts. We know that variables like the type of garment, price, the presence of shorts or skirts, customer age, and the denim trousers category play impactful roles in shaping consumer behavior within this category. In essence, these insights enhance our ability to harness the power of data and optimize e-commerce strategies and understandings effectively.

3.2.2 Testing

Moving onto the modeling aspect of the Random Forest, we used similar metrics as the ones used for the Lasso modeling. Similar to our Lasso classification, we are creating a balanced sample by doing a 50/50 split.

Table 2. Metric Evaluation for Random Forest

Random Forest	Average	Standard Deviation	Minimum	25th Quantile	Median	75th Quantile	Max
Accuracy	0.908048	0.002224	0.902606	0.906758	0.907939	0.909606	0.912970
Precision	0.957437	0.009672	0.921719	0.957731	0.959788	0.961373	0.966658
Recall	0.854154	0.010174	0.841587	0.849459	0.852056	0.854848	0.892082
R^2	0.632176	0.008889	0.610423	0.627021	0.631749	0.638400	0.651860
AUC	0.963391	0.001383	0.959620	0.962401	0.963306	0.964357	0.966150

According to the Random Forest model's results, it generally performs very well, with an accuracy of about 90.80%, in determining which category a product fall into. This indicates that it does a good job of classifying things into the appropriate groups—in this example, determining if an item is a garment for the lower body. Even more comforting is the model's small standard deviation of roughly 0.22%, which indicates that its performance does not significantly change between various scenarios. It performed poorly, reaching an accuracy of 90.26% at its lowest point and far above 91.30 at its greatest. However, it consistently stays in the range of 90.68% to 91.25%, with a median accuracy of 90.79%. We can conclude the Random Forest model is pretty consistent in its predictions and tends to stick around that range. In a nutshell, the Random Forest model is reliable and consistently performs well with high accuracy, making it incredibly useful for determining product categories within our dataset. for figuring out product categories in our dataset. The Random Forest had similar results when compared to the Lasso model.

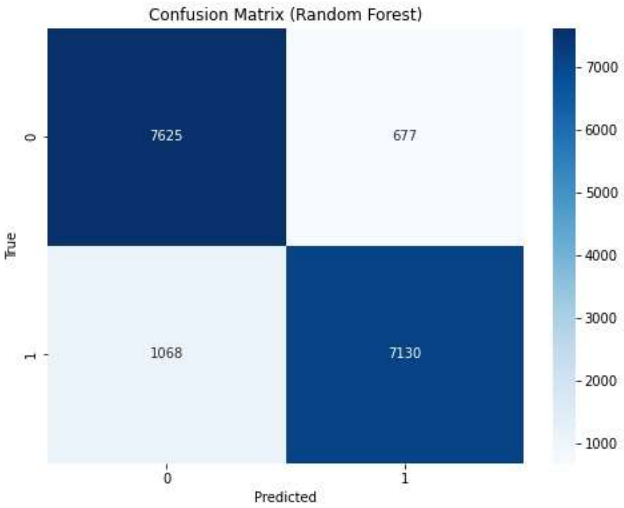

Figure 8. Confusion Matrix for Random Forest

Like in Lasso regularization, we do the same for our Random Forest classification model. That is, we displayed the true positives and true negatives of our classification models by creating a confusion matrix based on our findings. We took 33 percent of the values, or 16,500, out of the 50,000 total sample size for testing, since our confusion matrix provides a more thorough understanding of the specific areas where we made mistakes. We are able to see that we got 7258 true positives and 7672 true negatives. Similarly, we have 607 false positives and 963 false negatives. We can confidently say that our classification model is successfully making that distinction between lower body garment items (like pants, skirts, shorts) from other types of garments or items.

IV. Conclusion

The Random Forest model shows a similar performance to the Lasso model, with high accuracy, precision, recall, and (AUC). Both models appear to be strong candidates for our given task. The effectiveness of Random Forest and Lasso classification models in high-dimensional data contexts is highlighted by their comparable performance, which shows that both models display high levels of accuracy, precision, recall, and AUC.

These models exhibit consistent performance across multiple measures, indicating their robustness and dependability in categorizing an item. For e-commerce companies, our analysis can offer different applications that can benefit everyone. Businesses are better able to customize their inventory, marketing plans, and general customer experience when they comprehend the primary factors that influence consumer purchasing. Our models might also be used to improve search performance, automatically classify, and arrange things in online shopping, or even suggest outfit combinations to buyers. Correctly classifying clothing for the lower body could be useful in a variety of contexts outside of e-commerce, like retail store inventory management, recycling facility sorting, or fashion trend analysis.

4. 1 Future Work

In future research, we can aim to investigate other modeling approaches in order to improve our comprehension of consumer behavior and our capacity for prediction. Creating a neural network model is one direction we can plan to take. Neural networks have the ability to identify intricate patterns and connections within the data, which could lead to even more precise forecasts. We would also want to expand our analysis to include lower body garment purchase forecasts. Through the application of feature selection and clustering insights, predictive models that anticipate customer preferences and purchasing habits within this particular product category can be created. Businesses stand to gain a great deal from these forecasting capabilities since it will enable them to proactively adjust their lower body clothing inventory management and marketing tactics. Lastly, if we had more time, we would ideally aim to build a recommender system using more advanced machine learning techniques.

V. References

Accuracy vs. precision vs. recall in machine learning: What's the difference? Evidently AI - Open-Source ML Monitoring and Observability. (n.d.). https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

Bhadauriya, R. (2021, September 22). Lasso&nbsp;,Ridge &amp; Elastic net regression: A complete understanding (2021). Medium. https://medium.com/@creatrohit9/lasso-ridge-elastic-net-regression-a-complete-understanding-2021-b335d9e8ca3

Bhandari, A. (2023, August 31). Guide to AUC ROC curve in machine learning : What is specificity?. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Bootstrap aggregation, random forests and boosted trees. QuantStart. (n.d.). https://www.quantstart.com/articles/bootstrap-aggregation-random-forests-and-boosted-trees/

Carlos García Ling, ElizabethHMGroup, FridaRim, inversion, Jaime Ferrando, Maggie, neuraloverflow, xlsrln. (2022). H&M Personalized Fashion Recommendations. Kaggle; https://kaggle.com/competitions/h-and-m-personalized-fashion-recommendations

H&amp;M personalized fashion recommendations. Kaggle. (n.d.). https://kaggle.com/competitions/h-and-m-personalized-fashion-recommendations