Project Report: Fashion Retail Intelligence Project Report
Date: 11th August, 2025
To: Stakeholders
From: Shijin Ramesh, Data Analyst
---

1. Business Objective

Business Objective: The primary objective of this project is to leverage data analytics and machine learning to reduce return rates, and enable informed decision-making across various product categories and seasons.

2. Tools & Libraries

Used The analysis and model building were conducted using Python, along with several key libraries:
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Jupyter Notebook

3. Exploratory Data Analysis (EDA) Insights

Several key insights were gathered through exploratory data analysis:
- Return Rates by Product Category:
  - Bottoms have the highest return rate at 19.4%, followed by Shoes (17.5%) and Accessories (17.2%).
  - These categories may require quality checks, sizing reviews, or improved product descriptions to mitigate returns.
- Return Rates by Brand:
  - Ann Taylor exhibits the highest return rate at 15.9%, with Zara (15%) and Banana Republic (14.7%) close behind.
  - Given the strong reputation and quality of these brands, sizing issues are a likely reason for these unusual high return rates, potentially causing customers to return items.
- Return Rates by Season:
  - Summer recorded the highest return rate at 31.2%, followed by Winter (25.9%) and Spring (22.5%).
  - The spike in summer returns could be attributed to customers purchasing more items for the new academic year or vacations, leading to a higher frequency of size and style mismatches.
- Customer Ratings Analysis:
  - Shoes received the highest average customer rating (3.05), closely followed by Bottoms (3.02) and Dresses (3.01).
  - Outerwear, Accessories, and Tops had average ratings below 3, suggesting potential issues such as quality, sizing, or design preferences that warrant investigation.

◦ Overall customer ratings are concentrated between 3.0 and 3.5, with a noticeable second peak between 4.5 and 5.0, indicating some categories received excellent feedback.

◦ Customer ratings do not strongly depend on the product price, as high and low ratings are spread across all price ranges. This suggests that factors beyond price, such as quality, fit, or delivery experience, significantly influence customer satisfaction.

• Sales Trends:

◦ Sales were relatively stable for most months but spiked significantly in the latest month. This spike could be due to seasonal promotions, product launches, or other marketing activities that should be investigated for replication.

◦ Sales remain relatively consistent across all seasons, with only minor fluctuations.

◦ Highest sales are observed during the summer season, potentially due to parents purchasing new clothes for children at the start of a new academic year. Further analysis is recommended to confirm these driving factors.

◦ The Winter season shows the lowest sales, indicating potential opportunities for targeted promotions or seasonal marketing campaigns.

• Returned Items Volume:

◦ From August 2024 to July 2025, the number of returned items was relatively stable and low.

◦ However, in August 2025, returns suddenly spiked to 248 items, representing a sharp and unusual increase.

◦ This spike could be linked to end-of-season clearance sales or large promotional events in July, where customers often buy more discounted items but return them due to impulse purchases, incorrect sizing, or unmet product expectations.

4. Correlation Analysis

• A strong positive correlation exists between current price and original price, which is expected as current price is derived from the original price after discounts.

• There is also a clear negative correlation between current price and markdown percentage, meaning larger markdowns naturally lead to lower selling prices.

5. Model Building & Performance

• Objective and Approach: A classification model was built to predict whether a product would be returned (is_returned), as the target variable is binary.

• Features Used: The features selected for prediction include category, brand, season, size, original price, markdown percentage, stock quantity, customer rating, and purchase month.

• Initial Challenges with Decision Tree:

◦ The initial Decision Tree model achieved an accuracy of 74.54%, but it primarily predicted the majority class ('Not Returned').

◦ The model was able to predict only 4% of actually returned items, indicating an extremely imbalanced target variable (only 49 returned items in the test set).

◦ Due to class imbalance, the Decision Tree consistently leaned towards the majority class.

• Impact of SMOTE and Hypertuning:

◦ Applying SMOTE (Synthetic Minority Over-sampling Technique) helped the model identify more returned items, improving recall for True to 16% and precision for True to 14%. Accuracy dropped slightly.

◦ However, even after applying SMOTE and hypertuning the Decision Tree, the model's ability to identify returned items only slightly improved to about 19% recall, still misclassifying most returned items. This indicated that the Decision Tree, even with tuning, was not capturing strong patterns in the features for predicting returns, necessitating a stronger model.
• Transition to Random Forest: Based on the limitations of the Decision Tree, the project proceeded with a Random Forest model.

6. Key Findings from Model Results

Two machine learning models (Decision Tree and Random Forest) were tested to predict product returns based on attributes like season, size, brand, price, and customer ratings.
• Most Important Factors Influencing Returns: The most influential factors identified were season (Fall, Spring), product size (L), and brand (Ann Taylor), followed by original price and stock quantity.
• Model Limitations: Despite identifying these factors, both models were proficient at predicting non-returns but struggled significantly to accurately detect items that would be returned.
• Uncaptured Factors: This indicates that while some patterns are understood, returns are influenced by additional factors not captured in the current dataset. These may include customer behavior, delivery experience, or product quality issues post-purchase.

7. Recommendations to Stakeholders

Based on the insights and model findings, we recommend the following to stakeholders:
• Address High-Return Categories and Brands:
◦ For Bottoms, Shoes, and Accessories, implement rigorous quality checks, detailed sizing reviews, and enhance product descriptions to reduce their high return rates.
◦ For Ann Taylor, Zara, and Banana Republic, investigate potential sizing inconsistencies and consider implementing more detailed size guides or virtual try-on tools.
• Mitigate Seasonal Return Spikes:
◦ Investigate the specific reasons behind the significant spike in summer returns beyond academic year or vacation purchases.
◦ Consider pre-emptive measures during peak purchase seasons, such as improved sizing guidance or clearer product imagery, especially for items likely bought in bulk or for specific events.
• Leverage Customer Ratings:
◦ Investigate Outerwear, Accessories, and Tops categories that received lower average customer ratings to identify underlying issues related to quality, sizing, or design preferences.
◦ Since customer satisfaction is influenced by factors beyond price, focus on enhancing product quality, fit, and the overall delivery experience.
• Optimize Sales and Marketing Strategies:
◦ Analyze the significant sales spike in the latest month to understand its drivers (e.g., promotions, product launches) and identify strategies for replication.
◦ Despite summer having the highest sales, identify specific reasons for the lowest sales during Winter and explore opportunities for targeted promotions or seasonal marketing campaigns to boost performance.

◦ Investigate the sharp August 2025 returns spike (248 items) to confirm its link to end-of-season sales or promotional events and adjust return policies or product information during such periods.

• Improve Return Prediction Accuracy:

◦ To significantly improve the prediction accuracy of returns, we strongly recommend collecting more detailed data. This includes, but is not limited to:

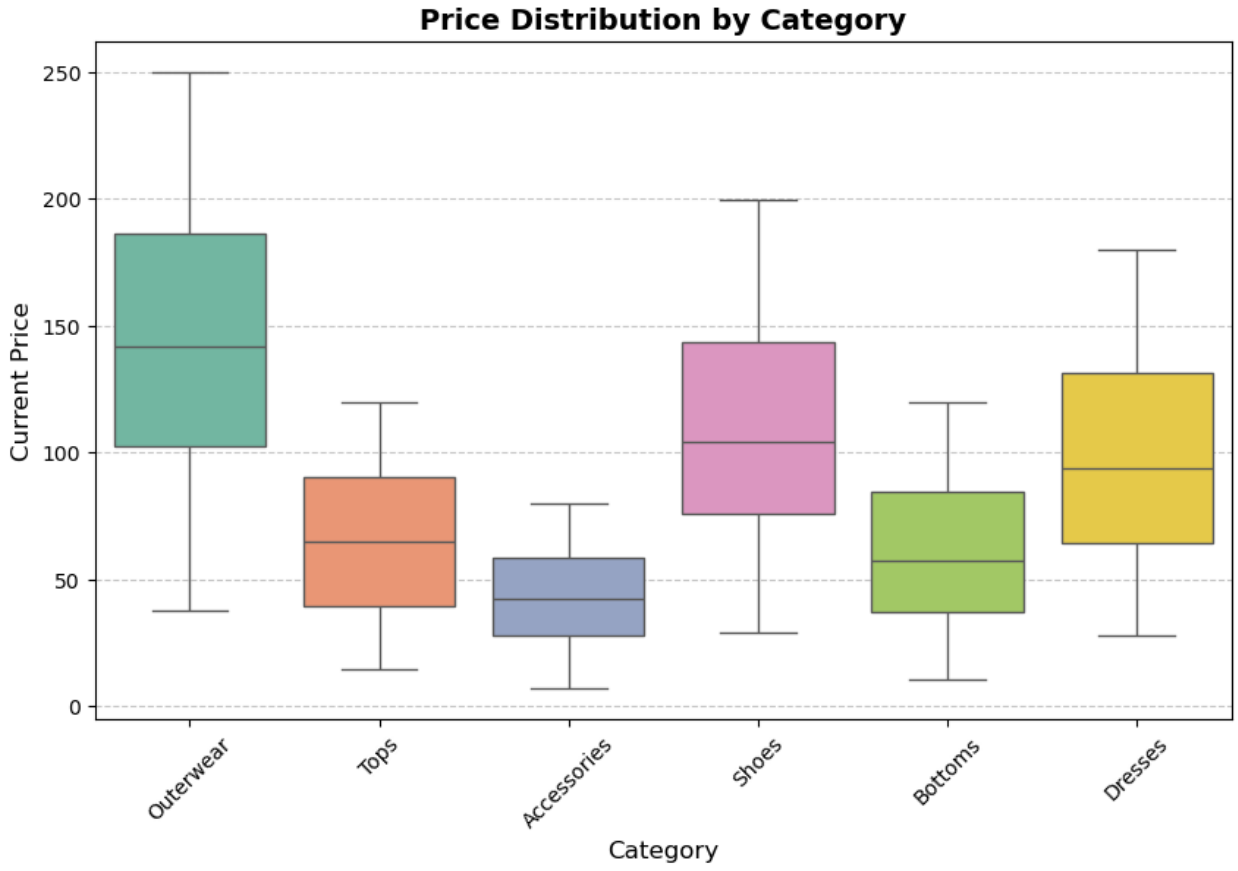
▪ Post-purchase customer feedback: Detailed reasons for returns, specific complaints.

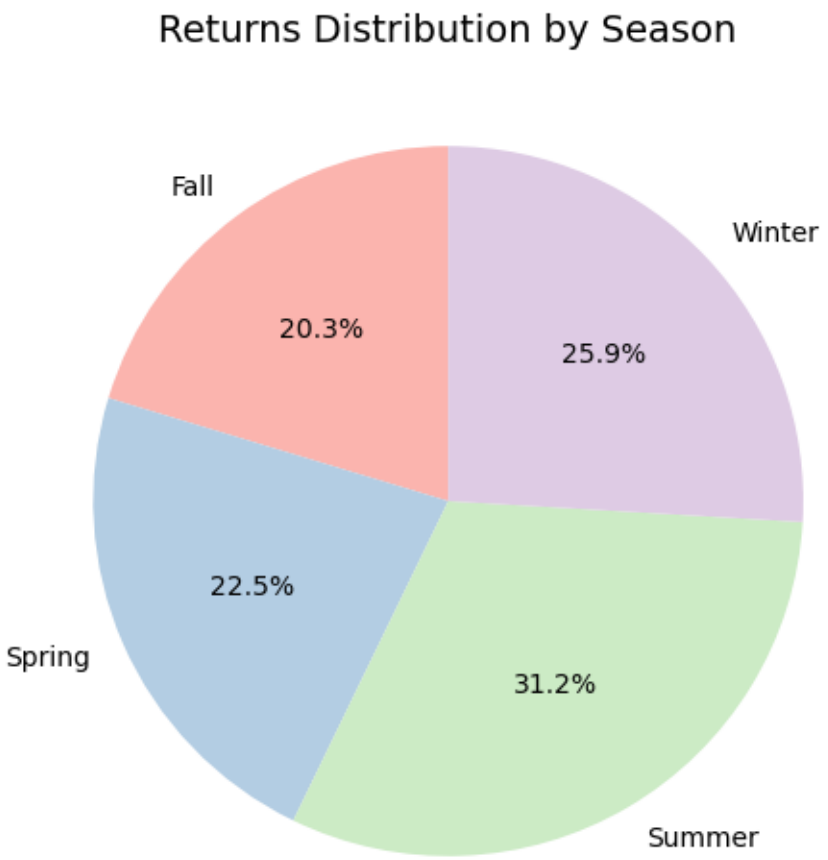
▪ Delivery timelines and experience: Information about shipping speed, packaging quality, and any delivery issues.

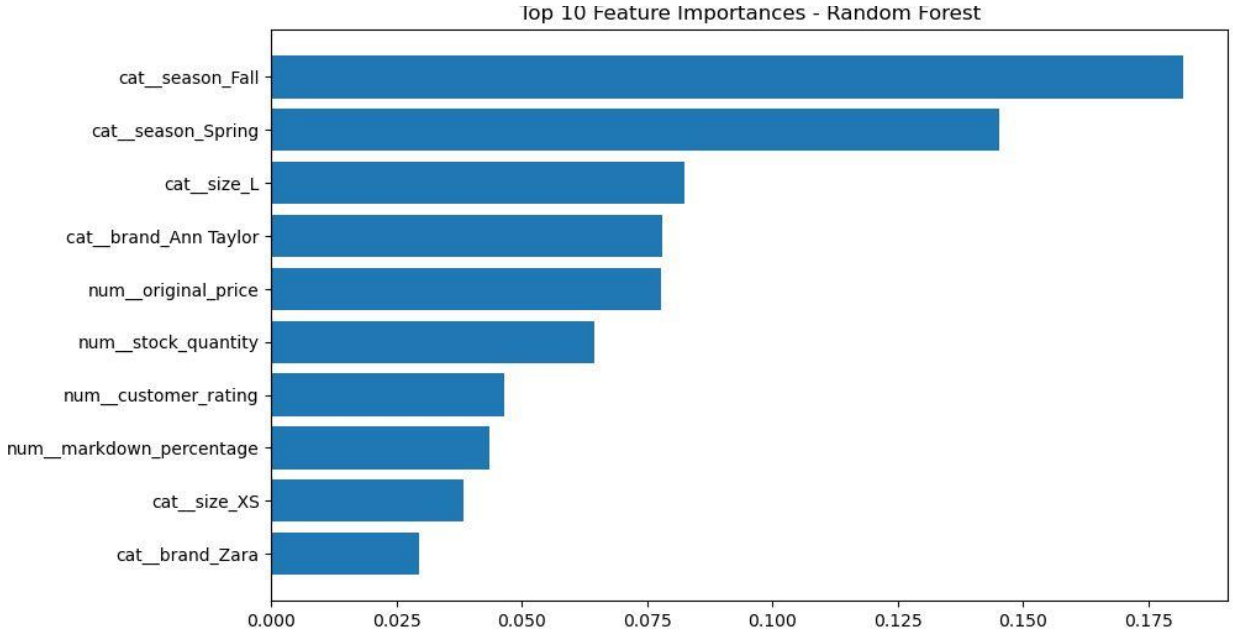
▪ Historical return behavior of individual customers: To understand recurring return patterns.

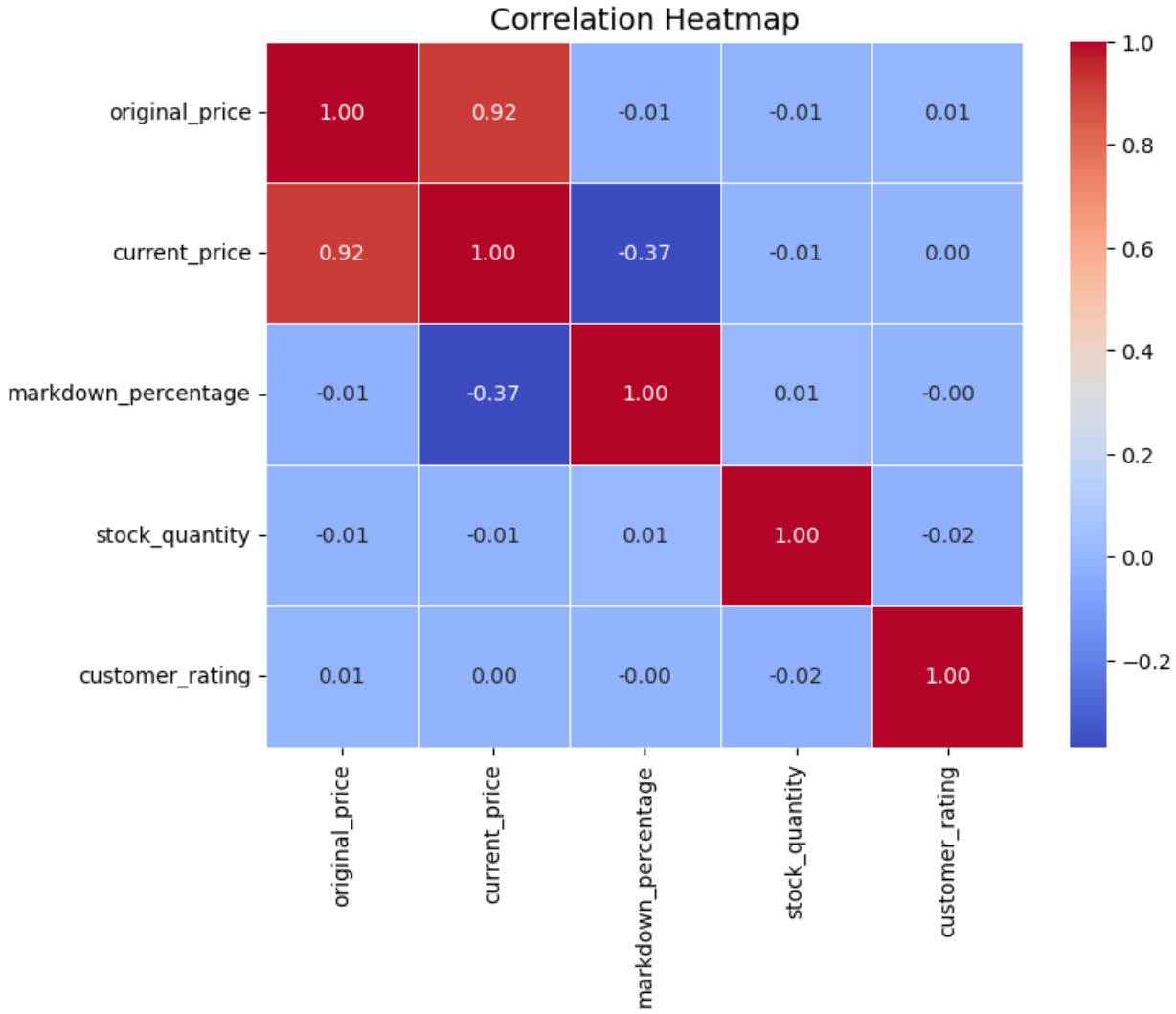
◦ This additional data will enable the development of a more robust machine learning model that is more balanced in predicting both returns and non-returns, moving beyond the current limitations.

8. Visualizations:




Returns Distribution by Season
Fall
Winter
20.3%
25.9%
22.5%
Spring
31.2%
Summer


Top 10 Feature Importances - Random Forest
cat__season_Fall
cat__season_Spring
cat__size_L
cat__brand_Ann Taylor
num__original_price
num__stock_quantity
num__customer_rating
num__markdown_percentage
cat__size_XS
cat__brand_Zara
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175


Correlation Heatmap
original_price
current_price
markdown_percentage
stock_quantity
customer_rating
1.00
0.92
-0.01
-0.01
0.01
0.92
1.00
-0.37
-0.01
0.00
-0.01
-0.37
1.00
0.01
-0.00
-0.01
-0.01
0.01
1.00
-0.02
0.01
0.00
-0.00
-0.02
1.00
original_price
current_price
markdown_percentage
stock_quantity
customer_rating
1.0
0.8
0.6
0.4
0.2
0.0
−0.2