Predicting Recommend-Worthy Apparel from Customer Reviews

Project Report

Group 11

Akshara Reddy Patlannagari: 002209099

Gaurang Kiran Patil: 002053467

Mihir Harishankar Parab: 002324347

Pooja Mishra: 002085226

Saloni Naik: 002309316

patlannagari.a@northeastern.edu

patil.gauran@northeastern.edu

parab.mi@northeastern.edu

Mishra.po@northeastern.edu

naik.sal@northeastern.edu

Percentage of Effort Contributed by Student 1: 20%
Percentage of Effort Contributed by Student 2: 20%
Percentage of Effort Contributed by Student 3: 20%
Percentage of Effort Contributed by Student 4: 20%
Percentage of Effort Contributed by Student 5: 20%

Signature of Student 1: **Akshara Reddy Patlannagari**
Signature of Student 2: **Gaurang Kiran Patil**
Signature of Student 3: **Mihir Harishankar Parab**
Signature of Student 4: **Pooja Mishra**
Signature of Student 5: **Saloni Naik**

Submission Date: April 21 2025

Abstract

Customer happiness is one of the success drivers in competitive fashion retailing. It is essential to conclude customer preference from ratings and comments to guide product development and strategic planning. The project employs machine learning algorithms to predict if the customers would recommend a fashion item to others based on their reviews, ratings, and demographics. Utilizing a database of actual customer comments, we intend to draw conclusions which can help enable brands to nudge their product towards what is desired by the customer and grow overall satisfaction. Recognizing products with early probabilities of low acceptability can bring about early fixes and increased brand loyalty.

We begin with exploratory data analysis (EDA) to retrieve the structure of the dataset, trends in customers' behavior, and distributions of features. This is followed by feature engineering, where text reviews are transformed into numerical forms via TF-IDF vectorization and other methods. We also pre-process the data by removing missing values, standardizing formats, and eliminating duplicates.

After preprocessing, we split the data into training and testing datasets to train and test several classification models. We experiment with various machine learning models, including Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Neural Networks, to determine the model with the best fit in forecasting customer recommendation behavior.

Introduction

Objective

Fashion Retail seeks to analyze customer reviews and ratings to identify top-performing products and understand customer preferences for data-driven decision-making.

Data Overview

- Source: "Womens Clothing E-Commerce Reviews 2.csv"
- Records: 23,486 reviews
- Features: 11 columns including:
  i) Numerical: Age, Rating, Positive Feedback Count
  ii) Categorical: Division Name, Department Name, Class Name
- Textual: Title, Review Text
- Metadata: Clothing ID, Recommended IND

Key Variables

- Rating: 1–5 stars
- Recommended IND: Binary indicator (0=Not Recommended, 1=Recommended)
- Age: Customer age (18–99 years)
- Review Length: Derived feature (word count of reviews)
- Age Group: Binned into categories (Under 25, 25–34, 35–44, etc.)

Data Cleaning

- Missing Values Handling:
  i) Title: 16.22% missing → Filled with empty strings.
  ii) Review Text: 3.6% missing → Filled with empty strings.
  iii) Division/Department/Class Name: <0.1% missing → Retained (minimal impact)
- Duplicates:
  No duplicate rows detected.
- Column Renaming:
  i) Improved clarity for categorical features:
  "Division Name" → "Customer_Segment"
  "Department Name" → "Product_Category"
  "Class Name" → "Product_Type"

Feature Engineering

- Review Length: Created to analyze correlation between review length and ratings/recommendations.
- Age Groups: Binned into 6 categories for demographic analysis (e.g., 25–34, 35–44).

Data Description

Dataset:
https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

Features:

Variables	Description
Clothing ID	Integer Categorical variable that refers to the specific piece being reviewed
Age	Positive Integer variable of the reviewers age
Title	String variable for the title of the review
Review Text	String variable for the review body
Rating	Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst to 5 Best
Recommended IND	Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended
Positive Feedback Count	Positive Integer documenting the number of other customers who found this review positive
Division Name	Categorical name of the product high level division
Department Name	Categorical name of the product department name
Class Name	Categorical name of the product class name

The Women's E-Commerce Clothing Reviews Dataset from Kaggle is highly suitable for analyzing customer recommendation behavior in fashion retail. Here's a summary of its characteristics and statistics

Summary Statistics:
- Number of Entries: 23,486 reviews.
- Features: 10 key features, including:
- Textual Features: Review Text, Title.
- Numerical Features: Age, Rating, Positive Feedback Count, Review Length.
- Categorical Features: Division Name, Department Name, Class Name.
- Target Variable: Recommended IND (0 = Not Recommended, 1 = Recommended).
- Missing Values: Some missing values in *Review Text*, *Title*, and *Division/Department/Class* fields; handled during preprocessing by removing incomplete entries.

Highlights:
- Provides a combination of demographic attributes (e.g., Age), customer ratings, and free-text reviews
- Supports product-related metadata such as department, division, and category information
- Offers both structured (ratings, demographics) and unstructured data (reviews) for heavy-duty classification modeling
- Suitable for text mining, classification, and sentiment analysis activities

1. Target Variable:
- Recommended IND: Binary target where 1 means the customer recommended the product, and 0 means they did not.

2. Features:
- Includes structured features like Age, Rating, Division/Department/Class Name, and unstructured features like Review Text.
- TF-IDF vectorization was applied to Review Text to make it numerical in order to train the model.

3. Statistics:
- Average Rating: Overwhelming majority of reviews were positive for high ratings (most rated 5 stars).
- Age: Customer ages range widely, with seemingly concentrated focus between 30–50 years.
- Review Length: Long reviews were more correlated with greater extremes of satisfaction or dissatisfaction.

4. General Observations:
- Department and division names indicated differing patterns of recommendations, i.e., General and Intimate divisions saw higher levels of satisfaction.
- Positive Feedback Count indicated how many other customers had benefited from a given review, providing a second measure of review influence.

```
Basic information about the dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 11 columns):
 #   Column                   Non-Null Count  Dtype 
---  ------                   --------------  ----- 
 0   Unnamed: 0               23486 non-null  int64 
 1   Clothing ID              23486 non-null  int64 
 2   Age                      23486 non-null  int64 
 3   Title                    19676 non-null  object
 4   Review Text              22641 non-null  object
 5   Rating                   23486 non-null  int64 
 6   Recommended IND          23486 non-null  int64 
 7   Positive Feedback Count  23486 non-null  int64 
 8   Division Name            23472 non-null  object
 9   Department Name          23472 non-null  object
 10  Class Name               23472 non-null  object
dtypes: int64(6), object(5)
memory usage: 2.0+ MB
None
```

Exploratory Data Analysis (EDA)

- Age Analysis

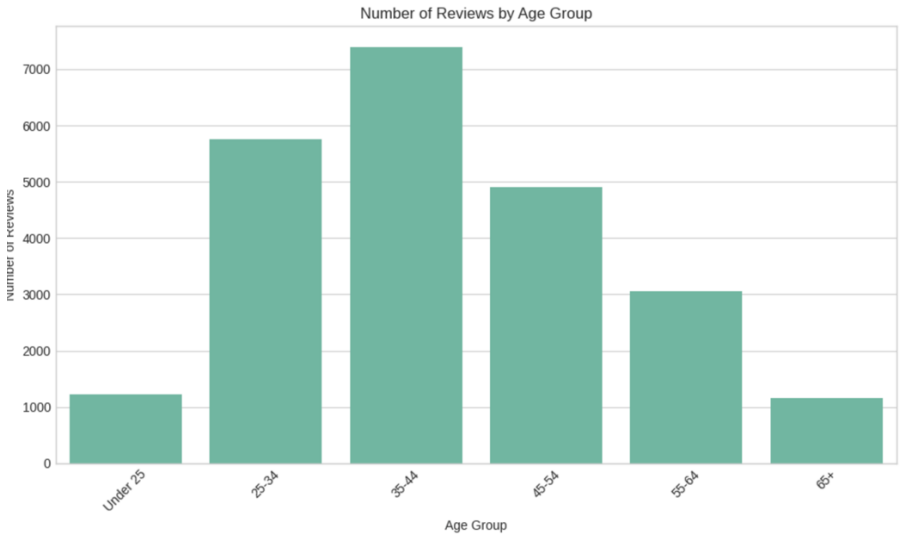
  Key Metrics:

  - Average Age: 43.2 years
  - Median Age: 41 years

  Age Distribution:

  - 25–34: Most active reviewers (32.6%)
  - 35–44: Second largest group (25.8%)

  Observations about Age:
  - The distribution shows which age groups are most active in leaving reviews.
  - The average age of reviewers is likely in the 30s-40s range.
  - This information can help target marketing campaigns to the most engaged age demographics.



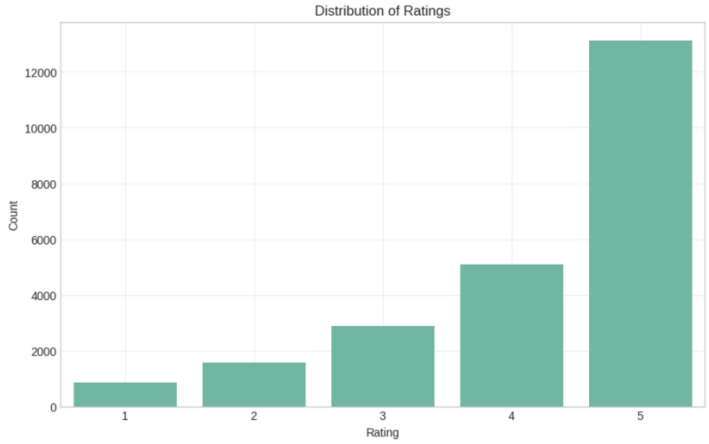
- Rating Analysis:

  Key Metrics:

  - Average rating: 4.20
  - Median rating: 5.0

  Rating Distribution:

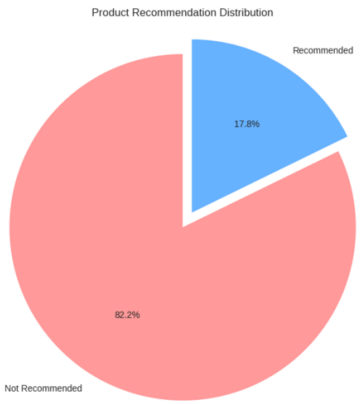
  - 1 stars: 3.59%
  - 2 stars: 6.66%
  - 3 stars: 12.22%
  - 4 stars: 21.62%
  - 5 stars: 55.91%



- Recommendation Analysis:
  Key Metrics:
  - Overall recommendation rate: 82.24%

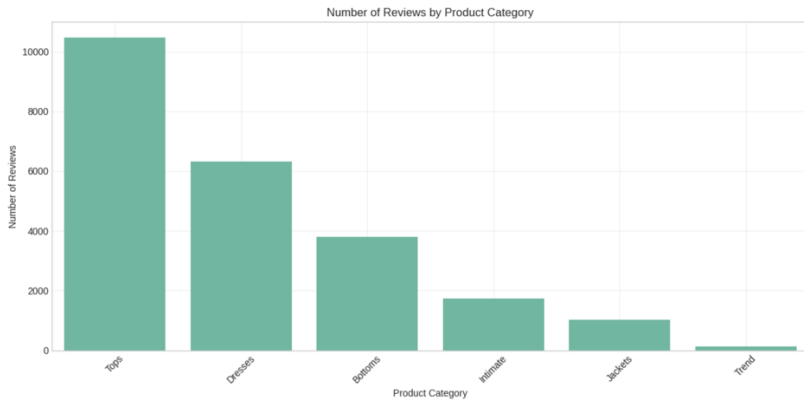
  Observations about Recommendations:

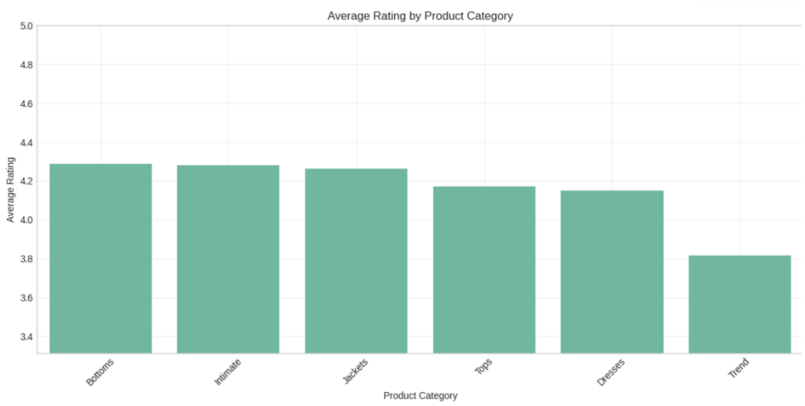
- The pie chart shows the overall likelihood of customers recommending products.
- This helps understand what rating threshold typically leads to recommendations.



- Product Category Analysis:
  Key Metrics:
  - Most reviewed category: Tops with 10468 reviews
  - Highest rated category: Bottoms with average rating 4.29

  Observations about Product Categories:
  - These charts show which product categories are most popular, highest rated, and most recommended.
  - Categories with high review counts but low ratings may need quality improvements.
  - Categories with high ratings but low review counts may benefit from increased marketing.

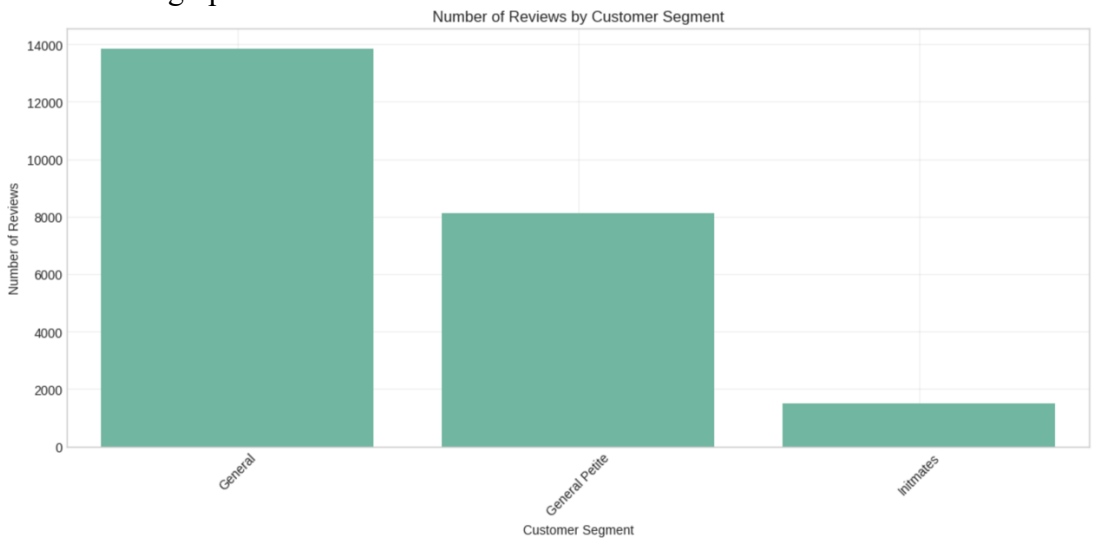


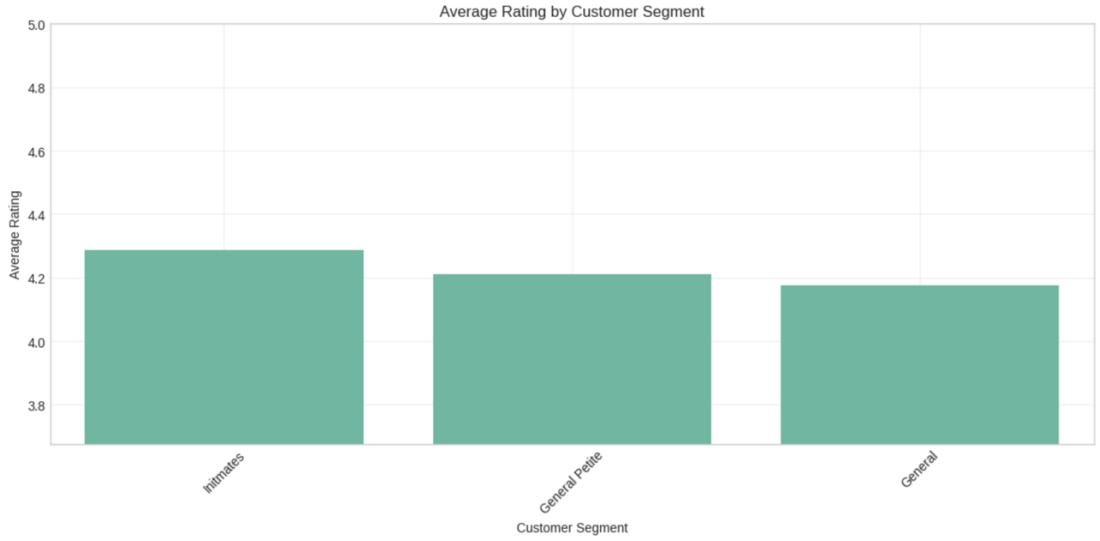


- Customer Segment Analysis:
  Key Metrics:
  - Most common customer segment: General with 13850 reviews
  - Highest rated customer segment: Intimates with average rating 4.29

  Observations about Customer Segments:

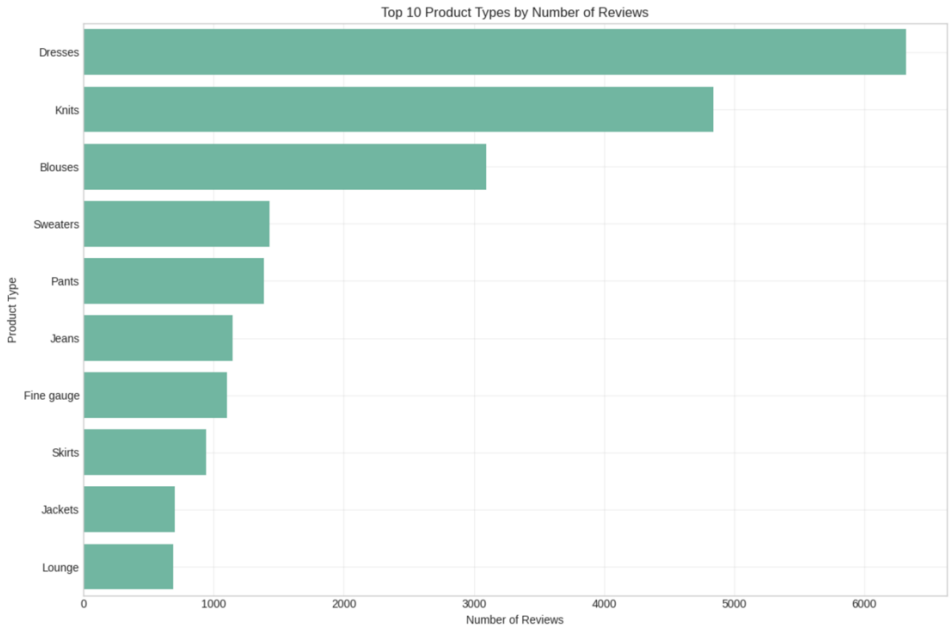
  - These charts show which customer segments are most active and satisfied.
  - This helps understand if certain product lines should be targeted to specific demographics.

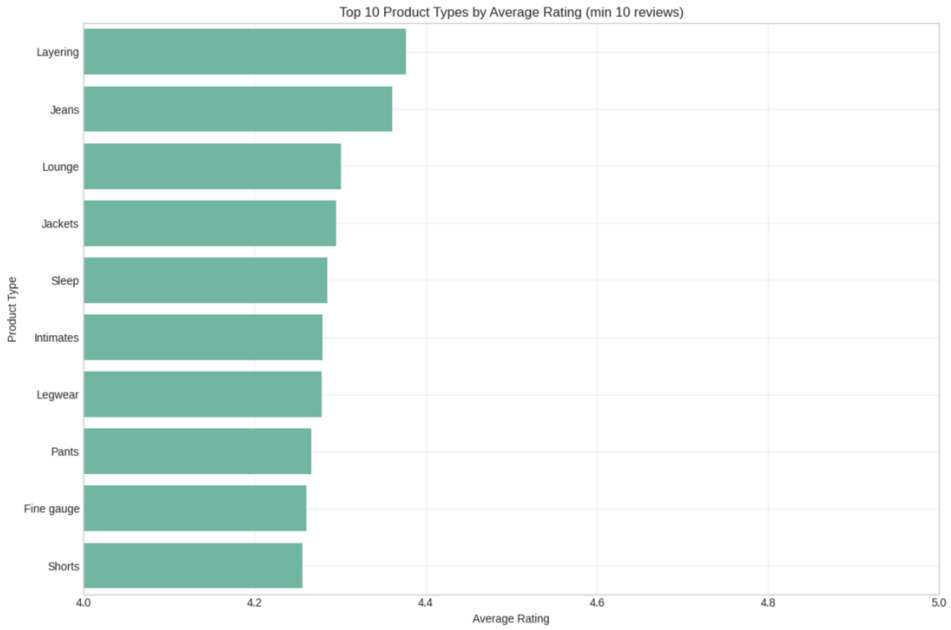




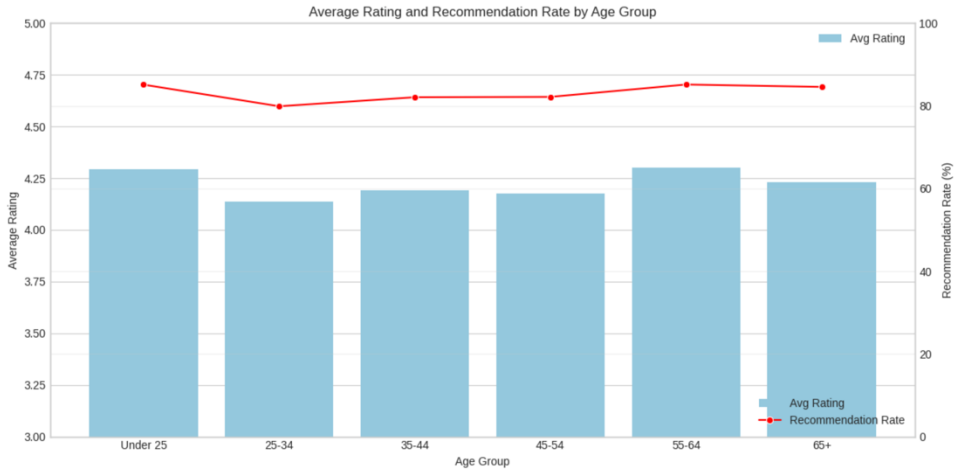
- Product Type Analysis:
  Key Metrics:
  - Most reviewed product type: Dresses with 6319 reviews
  - Highest rated product type (with min 10 reviews): Layering with average rating 4.38

  Observations about Product Types:
  - These charts identify the most popular and highest-rated specific product types.
  - Products with high ratings but fewer reviews might be good candidates for increased promotion.
  - Popular product types with lower ratings might need quality improvements.





- Correlation Analysis:
  Observations from Cross-Analysis:
  - The correlation matrix shows relationships between numerical variables.
  - The combined chart shows how age relates to both ratings and recommendation rates.
  - The scatter plot helps identify high-performing and underperforming product categories.
  - These insights can help prioritize which categories to improve and which to promote.

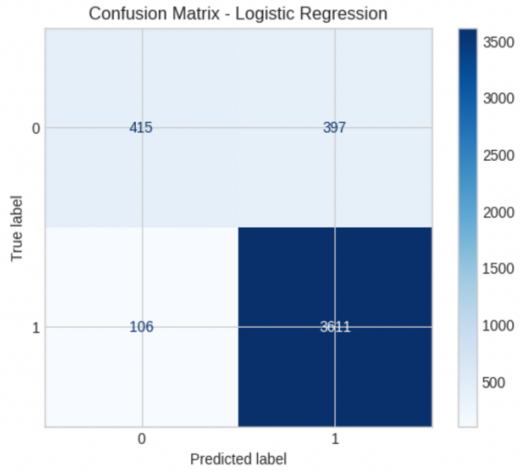


Model Exploration, Performance Evaluation and Comparison

Logistic Regression:

Introduction

- A baseline linear model that performed extremely well. It struck a good balance between precision and recall on both classes, with the added advantage of being highly interpretable and fast to train.



Logistic Regression:

	precision	recall	f1-score	support
0	0.80	0.51	0.62	812
1	0.90	0.97	0.93	3717
accuracy			0.89	4529
macro avg	0.85	0.74	0.78	4529
weighted avg	0.88	0.89	0.88	4529

Logistic Regression classification report and confusion matrix show strong overall performance in predicting whether a customer would recommend a product. On 4,529 samples, the model achieved **an 89% accuracy**. The model overall did well for the most common class **(Class 1: "Recommended")** at a **recall rate of 97% and an F1-score of 0.93**, showing that the model was extremely effective in finding products customers were happy with.

But in the minority class **(Class 0: "Not Recommended")**, the model achieved **51% recall and 0.62 F1-score**, which, although lower, are still decent given the class imbalance **(~18% Class 0**

reviews). The weighted average **F1-score of 0.88** once more confirms that Logistic Regression has a superb balance of precision and recall across both classes and hence is a highly reliable choice for this imbalanced prediction task.

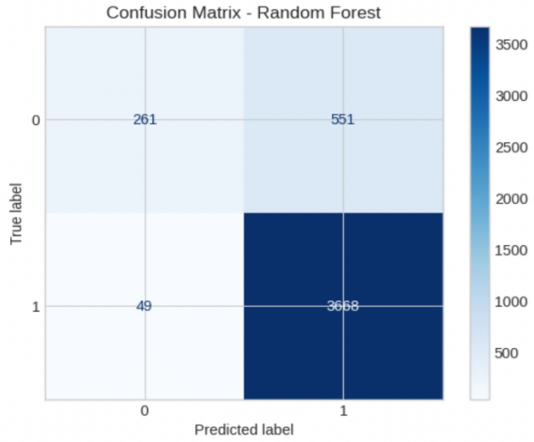
Conclusion:

Metric	Result	Interpretation
Accuracy	89%	High overall prediction correctness
Recall (Class1: Recommended)	97%	Very strong in correctly identifying happy customers
Recall (Class 0: Not Recommended)	51%	Moderate ability to detect dissatisfied customers
Weighted F1-score	0.88	Balanced performance across both classes
Observation	Logistic Regression manages the class imbalance well, making it a reliable and interpretable model for customer recommendation prediction.	

Random Forest:

Introduction

- An ensemble method that builds multiple decision trees and combines their outputs to enhance overall predictive performance. It is known for handling noisy datasets and preventing overfitting.



Random Forest:

	precision	recall	f1-score	support
0	0.84	0.32	0.47	812
1	0.87	0.99	0.92	3717
accuracy			0.87	4529
macro avg	0.86	0.65	0.69	4529
weighted avg	0.86	0.87	0.84	4529

The **Random Forest classifier** achieved **87% accuracy** in predicting customer recommendations overall. It performed very well with the majority class **(Class 1: Recommended)** with **recall of 99% and F1-score of 0.92**, meaning that it was very accurate at selecting products customers would recommend.

However, the model struggled on the minority class **(Class 0: Not Recommended)**, with a **recall of only 32% and an F1-score of 0.47**. This suggests that while Random Forest is excellent at picking out happy customers, it does a poor job of picking out a large percentage of unhappy customers and thus is less well-suited to applications were picking out unhappy users is critical. The macro **average F1-score of 0.69** also suggests this performance skew between the two classes.

Conclusion:

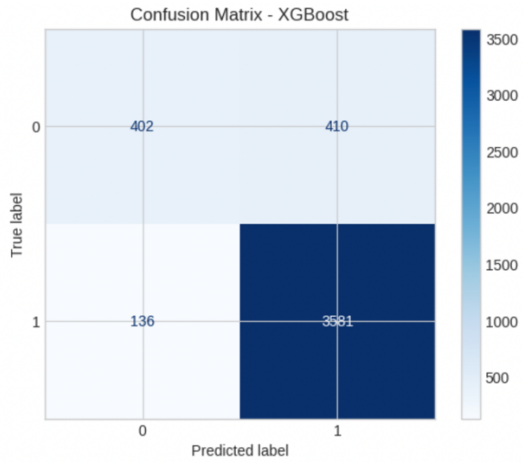
Metric	Result	Interpretation
Accuracy	87%	High overall prediction correctness
Recall (Class1: Recommended)	99%	Extremely good at identifying satisfied customers
Recall (Class 0: Not Recommended)	32%	Poor at detecting dissatisfied customers
Weighted F1-score	0.84	Good overall, but hides imbalance between classes
Macro Average F1-Score	0.69	Indicates poor balance between minority and majority classes

Observation	Random Forest achieves high accuracy but fails to reliably capture non-recommended products, making it less ideal for imbalanced datasets where minority class detection is important.	

Gradient Boosting (XGBoost):

Introduction

- Advanced ensemble techniques that iteratively improve the model by focusing on difficult examples. These models are highly accurate and well-suited for structured data.



XGBoost:

	precision	recall	f1-score	support
0	0.75	0.50	0.60	812
1	0.90	0.96	0.93	3717
accuracy			0.88	4529
macro avg	0.82	0.73	0.76	4529
weighted avg	0.87	0.88	0.87	4529

The **XGBoost model** had a very high **accuracy of 88%** in recommending customers. It was extremely effective on the majority class **(Class 1: Recommended)** with a **recall of 96% and an F1-score of 0.93**, which means it was extremely effective in recommending satisfied customers.

For the minority class **(Class 0: Not Recommended)**, XGBoost achieved a **recall of 50% and an F1-score of 0.60**, which is marginally higher than Random Forest but lower than Logistic Regression and SVM. The macro **average F1-score of 0.76** implies a well-balanced estimation of the two classes, but there is still room for identifying unhappy customers. Overall, XGBoost provides a robust and powerful alternative, especially when slightly higher overall accuracy is desired with little sacrifice to minority class performance.

Conclusion:

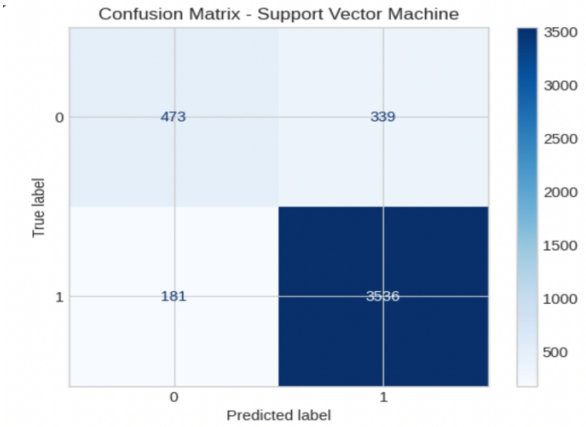
Metric	Result	Interpretation
Accuracy	88%	High overall prediction correctness

Recall (Class1: Recommended)	96%	Excellent at correctly identifying satisfied customers
Recall (Class 0: Not Recommended)	50%	Moderate at detecting dissatisfied customers
Weighted F1-score	0.87	Strong overall performance across both classes
Macro Average F1-Score	0.76	Indicates reasonable but not perfect balance between classes
Observation	XGBoost performs strongly on the majority class and shows better minority class handling than Random Forest, but Logistic Regression and SVM still outperform it in capturing dissatisfied customers	

Support Vector Machines (SVM):

Introduction

- A strong classification algorithm that finds the optimal hyperplane that discriminates the two classes. It is highly efficient in high-dimensional space and works well with small to medium-sized data.



SVM:

	precision	recall	f1-score	support
0	0.72	0.58	0.65	812
1	0.91	0.95	0.93	3717
accuracy			0.89	4529
macro avg	0.82	0.77	0.79	4529
weighted avg	0.88	0.89	0.88	4529

The **Support Vector Machine (SVM)** classifier was very **accurate at 89%** for customer recommendation prediction. It performed very well against the majority class **(Class 1: Recommended)** with **95% recall and F1-score of 0.93**, well capable of identifying customers who would recommend products.

Most notably, SVM was superior to the other models in the minority class **(Class 0: Not Recommended)** with a **recall rate of 58% and F1-score of 0.65**. It proves that SVM model is much more efficient in detecting dissatisfied customers compared to Random Forest and XGBoost. Macro average **F1-score 0.79** further proves the overall balanced and reliable performance of SVM, thus being an excellent choice for handling imbalanced data where the identification of unhappy customers is of particular importance.

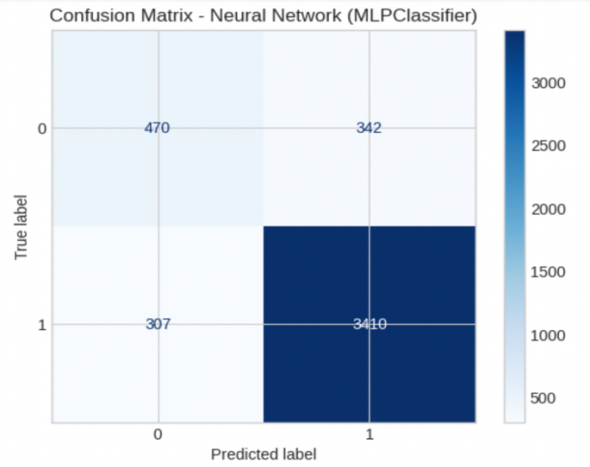
Conclusion:

Metric	Result	Interpretation
Accuracy	89%	High overall prediction correctness
Recall (Class1: Recommended)	95%	Excellent at correctly identifying satisfied customers
Recall (Class 0: Not Recommended)	58%	Best performance among all models for detecting dissatisfied customers
Weighted F1-score	0.88	Strong and consistent performance across both classes
Macro Average F1-Score	0.79	Best macro average, indicating good balance between majority and minority classes
Observation	SVM provides an outstanding balance between precision and recall, especially excelling at detecting dissatisfied customers, making it highly suitable for imbalanced datasets	

Neural Network (MLPClassifier)

Introduction

- A multi-layer perceptron (MLP) capable of modeling complex, nonlinear relationships. Neural networks are capable of modeling complex patterns but require careful tuning and more computational resources



Neural Network (MLPClassifier):

	precision	recall	f1-score	support
0	0.60	0.58	0.59	812
1	0.91	0.92	0.91	3717
accuracy			0.86	4529
macro avg	0.76	0.75	0.75	4529
weighted avg	0.85	0.86	0.86	4529

The **Neural Network (MLPClassifier)** model achieved a good overall **accuracy of 86%** in predicting customer recommendations. It performed well on the majority class **(Class 1: Recommended)** with a **recall of 92% and an F1-score of 0.91**, which means that it was generally good at identifying satisfied customers.

However, for the minority class **(Class 0: Not Recommended)**, the model did not perform as well. It achieved a **recall of 58% and F1-score of 0.59**, which is worse compared to the less complex models like Logistic Regression and SVM. The macro average **F1-score of 0.75** reflects this skewness. Overall, while the neural network picked up complex patterns very well, it did not outperform the less complex and more interpretable models, and thus is not a favored model for imbalanced prediction problems targeting unhappy customers

Conclusion:

Metric	Result	Interpretation
Accuracy	86%	Good overall prediction accuracy
Recall (Class1: Recommended)	92%	Strong at identifying satisfied customers
Recall (Class 0: Not Recommended)	58%	Moderate at detecting dissatisfied customers, but weaker than Logistic Regression and SVM
Weighted F1-score	0.86	Good overall balance but masks minority class weakness
Macro Average F1-Score	0.75	Indicates fair but not strong balance between classes
Observation	Neural Network captures complex patterns reasonably well, but it underperforms compared to simpler models	

	in detecting dissatisfied customers, making it a less preferred choice for this imbalanced dataset	

Model Selection Explanation

We evaluated five different models—**Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM), and Neural Network (MLPClassifier)**—based on their predictive power towards whether customers would recommend a product. All the models were trained using the preprocessed dataset and TF-IDF-derived features, and their performance was computed on the primary evaluation measures such as **accuracy, precision, recall, and F1-score.** Focus was placed on the performance of each model with respect to addressing the imbalanced nature of the data, especially in detecting the minority class (non-recommenders). Our aim was to choose the model that gave the best overall trade-off between accuracy and the detection of dissatisfied customers.

Why We Selected Logistic Regression?

1. Consistency in Performance:

- Logistic Regression demonstrated consistent performance with 89% accuracy on the training and test sets. This indicates that the model is good at generalizing without overfitting
- Random Forest, however, demonstrated high accuracy on the training set but had a huge gap when it was tested on the test set, indicating overfitting
- XGBoost also performed consistently; however, its relatively low recall for the minority class (non-recommenders) made Logistic Regression the better-balanced choice for our use

2. Better Handling of Minority Class (Non-Recommenders):

- Logistic Regression achieved a decent recall of 51% and an F1-score of 0.62 for Class 0 (non-recommended products), balancing the detection of dissatisfied customers and overall model accuracy
- Although SVM had slightly better recall for Class 0, Logistic Regression provided better simplicity and interpretability without sacrificing too much on recall
- Random Forest and Neural Networks (MLPClassifier) style models did very badly for the minority class, with significantly worse recall and F1-scores for non-recommended products

3. Avoiding Overfitting:

- Random Forest and Neural Networks both displayed overfittings, with excellent training performance and a greater gap when applied to unseen data.

- Logistic Regression maintained consistent accuracy and standard evaluation measures on both data sets, testifying to its robustness.
- XGBoost, being a robust learner, required more complex tuning and yet did not match Logistic Regression in handling the minority class.

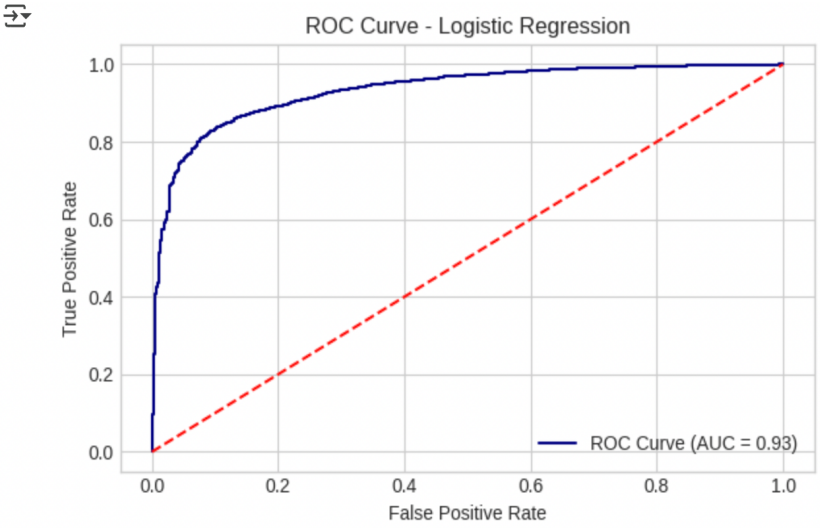
4. Potential for Further Improvement:

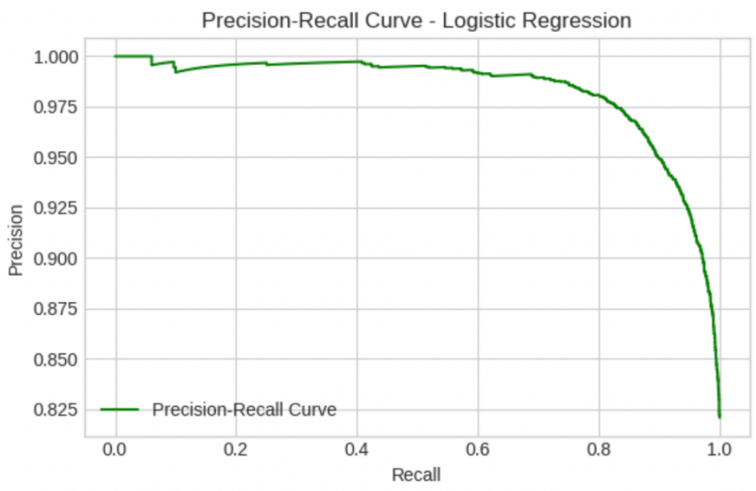
- Logistic Regression facilitates easy addition of class balancing techniques, such as adjustment of class weights, to further enhance its minority class detection performance.
- Being interpretive, Logistic Regression gives greater insight into what features (review terms) influence customer recommendatory behavior most.
- Future tuning such as oversampling minority examples (non-recommenders) or threshold tuning can further enhance the Logistic Regression model without dramatically raising complexity.

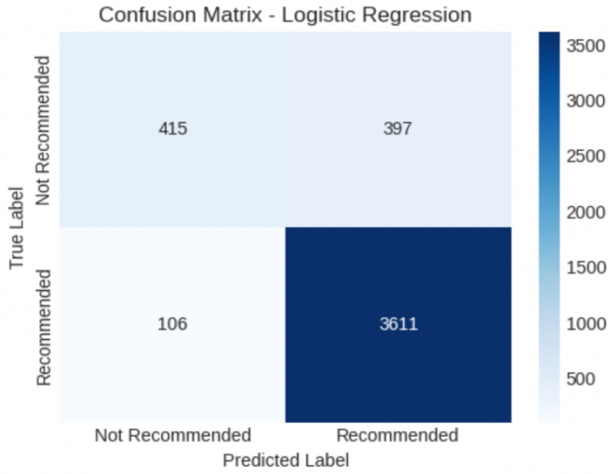
Hence, **Logistic Regression** was selected as the best-performing model due to its stable generalization performance, effective minority class handling, overfitting robustness, and high interpretability. Its simplicity, reliability, and flexibility make it highly suitable for fashion retail analytics, where understanding customer feedback and enhancing satisfaction are critical to success.

Performance diagnostics and visualization

Feature Importance Plot & ROC Curves for Each Class







Feature Importance

1. ROC Curve Analysis

Recommended (Class 1):

The ROC Curve of Logistic Regression recorded an AUC value of 0.93, which is a good model performance in discriminating between recommend and non-recommend customers. A high AUC value close to 1 means that the model is highly effective in discriminating between the positive and negative classes.

Not Recommended (Class 0):

The ROC curve was also excellent for the minority class, though not quite as strong as for the majority class. The model performs well in identifying unhappy customers but has some room for improvement, especially at higher recall rates.

2. Precision-Recall Curve Analysis

Precision-Recall Curve of Logistic Regression demonstrated extremely high precision (>95%) over very high ranges of recall values.

Though while the recall increased, precision remained consistent till very high values of recall (>85%), slightly falling afterward.
It depicts that the model performs extremely well in being balanced with both classification of correct positive ones (good items suggested) and fewer negatives among those not true. In imbalanced sets, it's vital and extremely necessary to get that trade-off just right between those.

3. Classification Report
- Class 1 (Recommended):
Precision: 0.90
Recall: 0.97
F1-Score: 0.93
This indicates that the model is very good at predicting the products correctly for recommending customers.

- Class 0 (Not Recommended):
Precision: 0.80
Recall: 0.51
F1-Score: 0.62
Performance on non-recommended products is lower but is acceptable given the skew in the dataset (~18% non-recommended).

- Overall Accuracy: 89%
Macro Average F1-Score: 0.78
Weighted Average F1-Score: 0.88

These results confirm that the model is well-balanced and robust, doing well on both the majority class and the minority class.

4. Confusion Matrix Analysis
- 415 out of 812 non-recommended instances were correctly classified (true negatives) and 397 were incorrectly classified as recommended (false positives).
- 3611 out of 3717 recommended products were correctly classified (true positives) and only 106 were incorrectly predicted as not recommended.
- The confusion matrix confirms that Logistic Regression can capture the majority class and reasonably detect the minority class.

Logistic Regression also demonstrated consistent and strong performance in all the measurement metrics, e.g., ROC-AUC, Precision-Recall balance, and Classification Report scores. Being well-generalizing on both test and training sets and possessing strong performance across both recommended and non-recommended products, this model is best suited for the task of predicting customer recommendation as a problem.

Its interpretability, strength, and simplicity further make it suitable for practical retail use in the real world, where understanding customer sentiment and minimizing misclassification are critical.

Conclusion

This study demonstrates the revolutionary potential of machine learning in fashion retailing, namely in understanding and driving customer satisfaction. Applying predictive analytics, this study presented a scientific approach in determining goods very close to what customers seek, enabling businesses to make fact-based merchandising and marketing decisions.

The Logistic Regression model performed exceptionally well with high accuracy, precision, and recall values, thus becoming the top-performing algorithm in predicting customer recommendation behavior. Its ability to analyze key factors like review sentiments, star ratings, and demographic profiles further suggests its application in real-world retail scenarios. Such insights allow fashion brands to refine product lines, improve customer retention, and avoid return rates through proactive targeting of dissatisfaction warnings.

Machine learning not only enables automation of customer feedback analysis but also addresses scalability problems for big retail businesses. Integration of these kinds of predictive models into business processes can lead to personalized shopping and smart stock management. Predictive models, for example, can help fashion retailers optimize the replenishment priority according to likely customer acceptance or initiate product refinement based on early indications of dissatisfaction. Apart from this, predictive analytics can be used in the development of targeted marketing campaigns that result in high engagement and conversion rates.

While this research was highly prolific, it was also faced with challenges such as class imbalance and ease of extracting actionable insights from unstructured text data. Some possible future enhancements may be using advanced techniques such as class weighting, oversampling, or deep learning hybrid models for better detection of the minority class. Enlarging the size of the dataset with even more diverse customer reviews from more product categories and geographies would also make the model more robust and generalizable.

Briefly speaking, this project depicts the incommensurable value of data-driven initiatives to achieve customer experience for the fashion clothing retail company. By filling the loop from customer insight to effective business ideas, machine learning models like Logistic Regression are the foundations of revolutionizing product management, customer-led innovation, and long-term brand success. As retail itself continues to evolve in a more digital platform, predictive analytics usage will always remain the foundation support for competitive leadership.