Predicting the 'Fit' of Clothing Items

Siddhant Bhagat
Department of Computer Science and Engineering
University of California San Diego
San Diego, California
sibhagat@ucsd.edu

Abstract

This report presents a detailed analysis and predictive modeling for predicting clothing fit ('small', 'fit', or 'large') for an online clothing rental service, using data from Rent the Runway. It begins with exploratory data analysis to understand customer preferences and sizing issues, followed by the development and comparison of three machine learning models: Logistic Regression, Naive Bayes, and Random Forest. The models were evaluated on accuracy, precision, recall, and F1-score, with an analysis of their confusion matrices. The report also discusses model optimization, scalability, overfitting, and alternative modeling approaches. The findings offer a comprehensive framework for predicting clothing fit and insights into customer sizing, crucial for improving user satisfaction in online fashion rental

1 Exploratory Analysis

1.1 Data Overview

The dataset contains **192,462 entries** with **15 columns**. Here's a brief description of each column:

Column Name	Description
fit	Describes how well the item fit the user.
user_id	Unique identifier for the user.
bust size	The bust size of the user.
item_id	Unique identifier for the item.
weight	Weight of the user.
rating	Rating given by the user.
rented for	The occasion for which the item was rented.
review_text	Text of the review given by the user.
body type	Body type of the user.
review_summary	Summary of the review.
category	Category of the item.
height	Height of the user.
size	Size of the item.
age	Age of the user.
review_date	Date of the review.

Table 1: Description of Columns in the Dataset

- **User Demographics:** The dataset predominantly features a younger female audience (based on the nature of the rental items and sizes). This is evident from the age and body measurements.
- **Ratings:** High ratings suggest customer satisfaction. However, the lack of low ratings could also indicate a bias in which users choose to leave reviews.
- **Size and Fit:** The wide range of sizes and the 'fit' variable indicate that the service caters to a diverse body size demographic.
- **Data Quality:** The presence of missing values and outliers (like age being 0 or 117) suggests the need for careful data cleaning, especially if this data is to be used for predictive modeling.

1.2 Data Cleaning

For data cleaning, I converted height to inches and weight to pounds for dataset. This ensured proper types of the columns. Age and Rating were string objects, they were created back to integer appropriately. The following is a table for the missing values in the dataset. The missing values were handled by either replacing them with 0 for numerical columns or by NaN for categorical columns

Column Name	Missing Values
fit	0
user_id	0
bust size	18397
item_id	0
weight	29957
rating	0
rented for	10
review_text	0
body type	14637
review_summary	0
category	0
height	677
size	0
age	959
review_date	0

Table 2: Count of Missing Values per Column

1.3 Descriptive Statistics

The following table shows some interesting stats for the numerical columns in the dataset. They give a good idea of the users and provide some insights of how might the fit be for a particular user

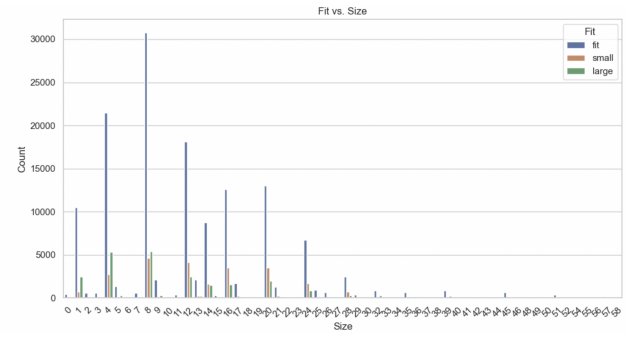
Column	Avg	Min	Max
Weight	137.39	50	300
Rating	9.09	2	10
Height	65.31	54	78
Size	12.25	0	58
Age	33.87	0	117

Table 3: Descriptive Statistics Summary

1.4 Data Visualization

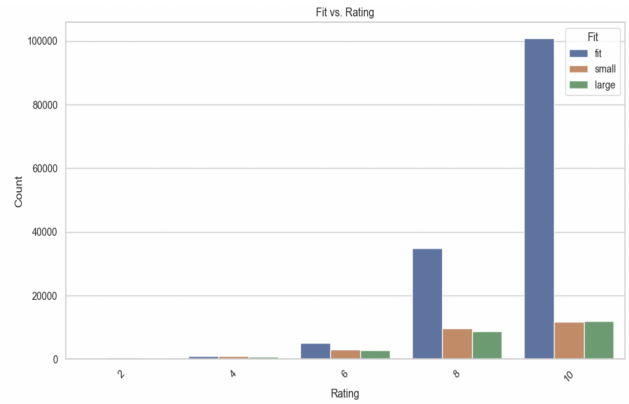
Fit vs. Size:

This plot shows the distribution of 'fit' across different sizes. ¡br¿ There is a noticeable trend where certain sizes have a higher proportion of 'fit' outcomes, while others have more instances of 'small' or 'large'. This suggests that the fit issue may be more prevalent in specific size ranges.



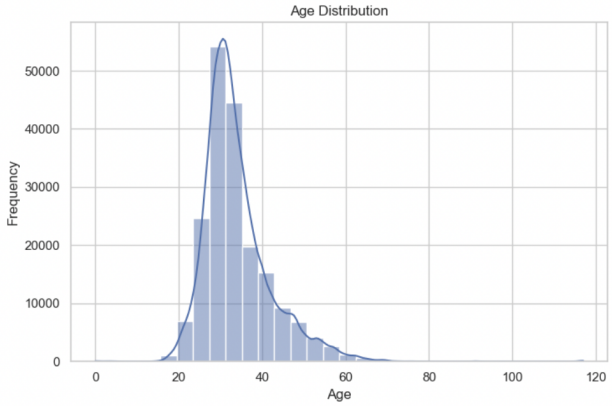
Fit vs. Rating:

The relationship between fit and rating is evident in this plot. ¡br¿ Ratings tend to be higher when the fit is good. Conversely, when the fit is 'small' or 'large', the ratings appear to decrease, albeit not drastically. This indicates that while fit impacts customer satisfaction, other factors also play a significant role in determining the overall rating.



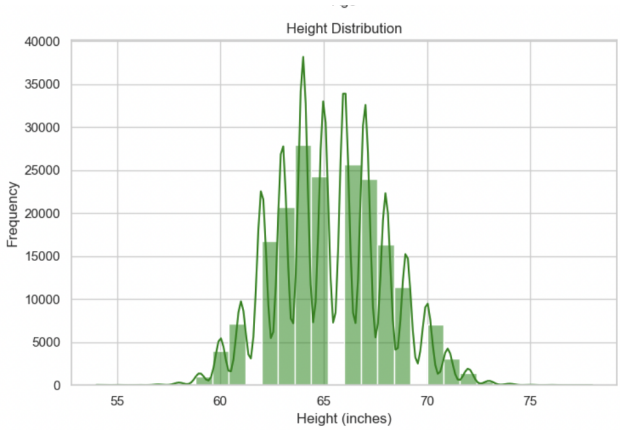
Age Distribution:

The majority of users are in their late 20s to early 40s. The distribution is slightly right-skewed, indicating a younger user base.



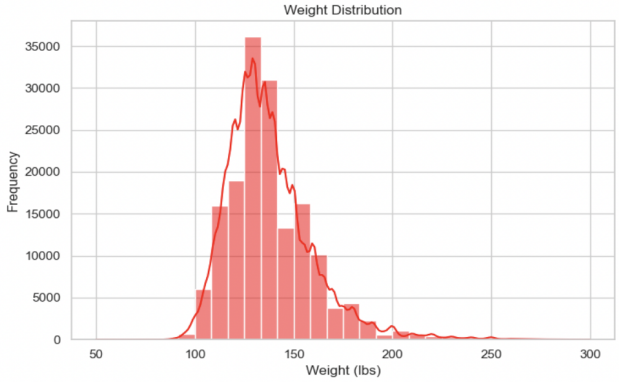
Height Distribution:

The height of users seems normally distributed. Most users are around 63 to 67 inches tall (about 5'3" to 5'7").



Weight Distribution:

The weight distribution is roughly normal but slightly right-skewed. Most users weigh between 120 and 150 lbs.



2 Predictive Task

2.1

Defining the Predictive Task

I am setting out to predict the fit of clothing items (categorized as 'small', 'fit', or 'large') for users in the dataset. This is a classification problem where I aim to match users with the right fit based on their attributes and the characteristics of the clothing.

2.2

Baseline Models

1. Simple Logistic Regression: Using `LogisticRegression()` from Scikit-Learn with default settings. This model serves as a basic yet often effective benchmark for classification tasks.
2. Another Simple Baseline: I will choose a Naive Bayes classifier as the second baseline, considering its simplicity and effectiveness in classification problems, particularly as a starting point.

2.3 Features to Use

I will focus on features that are likely to influence the fit, such as:

1. User-related features: 'age', 'height', 'weight', 'body type'.
2. Clothing-related features: 'size', 'category', 'rented for'.
3. Derived features: For instance, Body Mass Index (BMI) calculated from 'height' and 'weight'.

2.4 Metrics

To evaluate the performance of the predictive model for clothing fit, the following metrics will be used:

1. **Accuracy**: Measures the overall correctness of the model.
   - Equation:
   $$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$
2. **Precision (for each class 'small', 'fit', 'large')**: Indicates the proportion of positive identifications that were actually correct.
   - Equation:
   $$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
3. **Recall (for each class 'small', 'fit', 'large')**: Measures the proportion of actual positives that were identified correctly.
   - Equation:
   $$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$
4. **F1-Score (for each class 'small', 'fit', 'large')**: Provides a balance between Precision and Recall.
   - Equation:
   $$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.5 Validity Assessment

To ensure the validity of my model's predictions, I will:

1. Use Cross-Validation: To evaluate the model's performance across different subsets of the data, ensuring that the model is not overfitting to a particular part of the dataset.
2. Compare with Baseline Models: Evaluate how much better my model performs compared to the simple Logistic Regression and Naive Bayes baseline models.
3. Error Analysis: Examine the types of errors made by the model (e.g., confusing 'small' with 'large') to understand its limitations and potential areas for improvement.
4. Test Set Performance: Assess the model's performance on unseen data, which is crucial for evaluating its generalizability.

2.6 Relevant Tables Using Data

Class Distribution of Fit: This table shows that the majority of the entries are classified as 'fit', with 'small' and 'large' being less frequent. This distribution is important for understanding the performance of baseline models and the potential need for handling class imbalance.

Fit Category	Count	Proportion
fit	141995	0.737782
small	25776	0.133928
large	24691	0.128290

Fit Across Categories: The fit across categories table can help identify if certain clothing types tend to have more fit issues. For example, categories with a high number of 'small' or 'large' fits relative to 'fit' might indicate sizing inconsistencies in those categories. Only the first few rows are shown.

Category	Fit	Large	Small
ballgown	10.0	5.0	1.0
blazer	595.0	117.0	70.0
blouse	484.0	107.0	60.0
blouson	11.0	2.0	1.0
bomber	94.0	14.0	20.0

Fit Across Body Types: The fit across body types table provides insights into how well the clothing items cater to different body shapes. Body types with a higher proportion of 'small' or 'large' fits might indicate a need for better sizing adjustments for those groups. First few rows:

Body Type	Fit	Large	Small
apple	3487	665	725
athletic	32444	5248	5960
full bust	10769	2092	2138
hourglass	40823	7187	7298
pear	16130	3189	2807

3 Model

3.1 Baseline 1: Logistic Regression

In our first baseline approach, I employed **Logistic Regression** to predict the fit of clothing items. Logistic Regression, known for its simplicity and interpretability, serves as an ideal baseline model. It's particularly advantageous due to its computational efficiency, which is crucial for initial assessments. However, it assumes a linear relationship between independent variables and the dependent variable, which might not always hold true, especially in scenarios with complex relationships.

The model was trained on a dataset encoded using `LabelEncoder` for categorical features like *'body type'*, *'category'*, and *'rented for'*. This transformation is crucial for converting categorical text data into a model-understandable numerical format. Our feature set included variables such as *'size'*, *'age'*, *'height'*, *'weight'*, *'body type'*, *'category'*, and *'rented for'*, which are intuitive predictors for clothing fit.

I split the data into training and testing sets, with 70% of the data used for training and the remaining 30% for testing, ensuring a robust evaluation of the model's performance. Prior to fitting the model, I standardized the features using `StandardScaler`. This step is critical for normalizing the range of independent variables, which helps in speeding up the convergence of the logistic regression algorithm.

Upon evaluating the model, it achieved an accuracy of 73.56%, a precision of 58.80%, a recall of 73.56%, and an F1-score of 62.56%. The relatively high recall indicates the model's effectiveness in identifying correct fits, but the lower precision suggests some challenges in accurately classifying the fit of clothing items, possibly due to class imbalances or the inherent simplicity of the model.

Given these metrics, the Logistic Regression model serves as a good starting point, highlighting the potential complexity of the dataset and paving the way for more sophisticated models or feature engineering to improve prediction accuracy.

Metric	Value (%)
Accuracy	73.56
Precision	58.80
Recall	73.56
F1-Score	62.56

Table 4: Performance Metrics of Logistic Regression Model

3.2 Baseline 2: Naive Bayes

In our exploration of baseline models, I next employed the **Naive Bayes Classifier**. This model is renowned for its effectiveness in handling categorical data, attributing to its suitability for our classification task. The Naive Bayes model is particularly advantageous due to its probabilistic approach, making it adept at managing uncertainty, and its efficiency, both in terms of computational resources and ease of implementation. However, it is limited by the naive assumption of feature independence, which might not reflect the complexity of real-world data. Additionally, while Gaussian Naive Bayes is apt for continuous data, it presumes a normal distribution, which is not always the case.

The implementation of this model followed a similar procedure as with the Logistic Regression model. I used the Gaussian Naive Bayes model from Scikit-Learn, considering the continuous nature of some of our features. The model's performance was evaluated using the same metrics as before: accuracy, precision, recall, and F1-score. The data was similarly split into training and testing sets, with 70% of the data used for training and the remaining 30% for testing

The Naive Bayes model demonstrated a similar accuracy level to the Logistic Regression model at 73.58%. However, its precision was slightly lower at 58.01%, hinting at potential challenges in classifying classes accurately. The recall and F1-score, at 73.58% and 62.62% respectively, are comparable to those of the Logistic Regression model, indicating a balanced performance in terms of precision and recall, yet with an evident need for improvement.

This performance suggests that while the Naive Bayes classifier can identify certain patterns within the data, its assumption of feature independence might limit its ability to accurately capture more intricate relationships. As with the Logistic Regression model, the insights gained from the Naive Bayes classifier highlight the potential need for more sophisticated modeling techniques or feature engineering to enhance the prediction of clothing fit.

Metric	Value (%)
Accuracy	73.58
Precision	58.01
Recall	73.58
F1-Score	62.62

Table 5: Performance Metrics of Naive Bayes Classifier

3.3 Random Forest Classifier

In advancing our model sophistication, we explored the **Random Forest Classifier**. This model is particularly adept at handling complex, non-linear relationships between features, making it a significant step up from Logistic Regression and Naive Bayes. Its ensemble nature contributes to robustness against overfitting, especially beneficial in high-dimensional datasets. Moreover, Random Forest elucidates feature importance, offering valuable insights into the factors most influencing the 'fit' prediction. It also handles imbalanced data effectively and works well with both numerical and categorical features, without the need for data scaling.

Despite showing a slightly lower accuracy of 70.58% compared to the baseline models, Random Forest exhibited a higher precision of 62.20% and an F1-score of 64.43%. This indicates a better balance between precision and recall, crucial for a more nuanced understanding of model performance.

The strengths of the Random Forest model lie in its increased precision and F1-score, suggesting enhanced handling of the trade-off between incorrectly predicting a class and missing a correct prediction. The model's capacity to capture more complex data relationships contributes to these improved metrics. However, the slightly lower accuracy

might stem from the model's complexity and the inherent randomness of the ensemble method. This suggests potential areas for improvement through feature engineering or parameter tuning.

In conclusion, while the Random Forest model does not substantially outperform the simpler models in terms of accuracy, it offers a more balanced performance between precision and recall. Its ability to manage complex data relationships is a noteworthy advantage. However, considerations of its complexity and computational demands are crucial for further model development and optimization.

Metric	Value (%)
Accuracy	70.58
Precision	62.20
Recall	70.58
F1-Score	64.43

Table 6: Performance Metrics of Random Forest Classifier

Parameter Tuning

In the process of optimizing the Random Forest Classifier, parameter tuning plays a crucial role. It involves adjusting various hyperparameters of the model to improve its performance. For our model, two key parameters were focused on: **n_estimators** and **max_depth**.

The *n_estimators* parameter determines the number of trees in the forest, and *max_depth* controls the maximum depth of each tree. By varying these parameters, we aim to find a balance between the model's ability to capture complex patterns and its generalization to unseen data. A higher number of estimators can potentially improve the model's accuracy but at the cost of increased computational complexity. Similarly, a deeper tree might capture more detailed data patterns, but it also risks overfitting.

The following table summarizes the performance metrics of the Random Forest model for different combinations of these two parameters:

n_estimators	max_depth	Acc	Prec	Recall	F1
10	None	0.6895	0.6120	0.6895	0.6379
30	2	0.7368	0.5429	0.7368	0.6251
50	5	0.7368	0.5429	0.7368	0.6251
70	8	0.7369	0.6321	0.7369	0.6255
80	10	0.7373	0.6427	0.7373	0.6267
100	11	0.7378	0.6970	0.7378	0.6282
120	12	0.7378	0.6785	0.7378	0.6289

Table 7: Performance Metrics of Random Forest Classifier with Different Parameter Settings

It can be seen that at the end of the seventh iteration of parameter tuning, the metric were not increasing substanitially, so I decided to stop there and report the results

Side Notes

- **Issues:**
  - Scalability can be a concern due to its computational intensity, especially with large datasets.
  - There's also a risk of overfitting if the trees are too deep.
- **Other Models Considered:**
  - Models like Support Vector Machines (SVM), Gradient Boosting Machines (GBM), and neural networks were considered for their ability to capture complex, non-linear relationships.
- **Unsuccessful Attempts:**
  - Early attempts included using simpler models without adequate feature preprocessing, leading to suboptimal performance.
  - Initial hyperparameter settings for complex models like Random Forest were not ideal, requiring iterative tuning.
  - Overfitting was an issue, especially in models like Random Forest, if the depth of the trees was not controlled properly.

4 Literature

The dataset used in this report was downloaded from Professor Julian McAuley's archive of datasets, which can be found here

The first piece of literature that I came across while finding similar studies that exist on predicting the 'fit' of clothing items was ***A novel approach in predicting virtual garment fitting sizes with psychographic characteristics and 3D body measurements using artificial neural network and visualizing fitted bodies using generative adversarial network*** by *Nga Yin Dik, Paul Wai Kei Tsang, Ah Pun Chan, Chris K.Y. Lo, and Wai Ching Chu*. The aim was to develop a virtual garment fitting prediction model that considers both body dimensions and psychographic characteristics. This model aims to address the challenge of defining ease allowances in virtual garment fitting for user satisfaction.

The data collection process involved 3D body measurements, psychological surveys, and garment fit assessments from subjects. Moreover, physical and psychological data was also collected. To understand the correlation between body measurements, garment fit, and consumer's psychographic profiles. There were two main methods utilized in this study:

1. **Artificial Neural Network (ANN)**: Employed to predict ease preferences based on body dimensions and psychographic characteristics.
2. **Generative Adversarial Network (GAN)**: Used to visualize fitted bodies in 3D with the predicted pattern parameters from the ANN model.

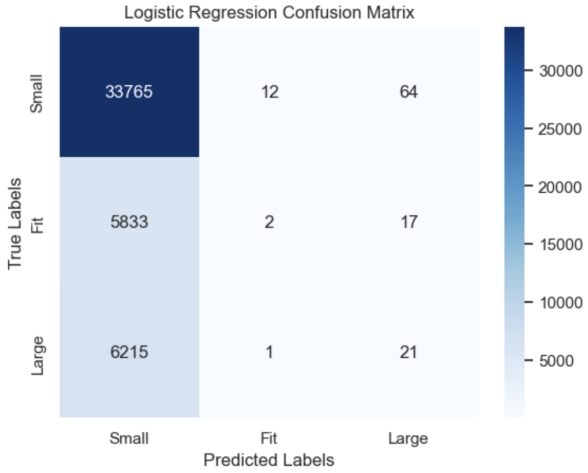
The study successfully developed an Artificial Neural Network (ANN) model that effectively predicts garment fit preferences. This approach revealed a nuanced, non-linear relationship between these factors and garment fitting parameters. Additionally, the utilization of a Generative Adversarial Network (GAN) for visualizing 3D fitted garments further validated the model's effectiveness. The results demonstrated potential for customization in the apparel industry, leading to user satisfaction. he study not

only generated new, more accurate size-fitting data but also highlighted the importance of considering psychological factors alongside physical measurements in garment design.

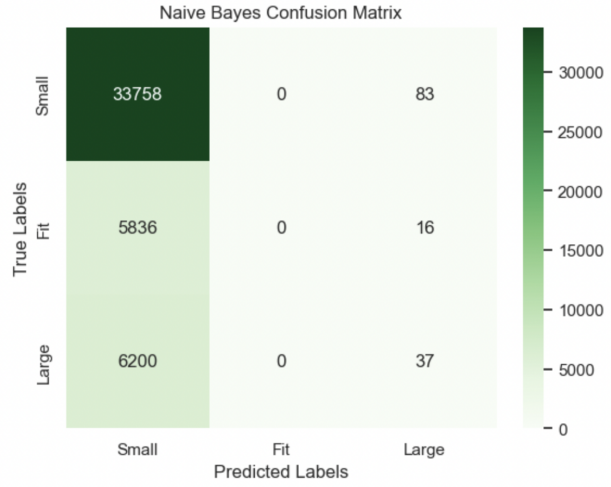
The second piece of literature that I came across was ***A Deep Learning System for Predicting Size and Fit in Fashion E-Commerce.*** The aim of the study is to develop a deep learning-based methodology for personalized size and fit prediction in fashion e-commerce. This approach addresses the challenge of size-related returns and customer satisfaction by leveraging both interaction data and customer/article features. The paper proposes a novel method that uses a split-input neural network architecture, capable of capturing both global and entity-specific properties for personalized recommendations.
The methodology, termed "Size and Fit Network" (SFnet), employs a non-identical feedforward input pathway architecture, capable of learning personalized latent features of individual customers and articles. This architecture allows the model to capture latent information about both entities contained in implicit patterns in the data, enabling it to identify multiple personas or intrinsic properties of certain articles or brands. The method's performance was compared with several methodologies, using metrics like the area under the ROC curve (AUC), accuracy, and average log-likelihood. This method produced outstanding results in predicting sizes of clothing items.
In conclusion, the study successfully demonstrates the efficacy of the proposed deep learning methodology in providing personalized size and fit recommendations. It highlights the potential of this approach to enhance customer satisfaction and reduce costs related to size-related returns in the fashion e-commerce industry.
The second paper, while also employing deep learning techniques, focuses more on a content-collaborative approach for size and fit prediction in e-commerce. It leverages both customer-article interaction data and additional customer/article attributes, addressing the sparsity problem common in collaborative filtering. The second paper's methodology is noteworthy for its practical application in e-commerce platforms, demonstrating improvements in predicting correct sizes and thus reducing return rates. While both studies show advancements in using deep learning for fashion-related challenges, the first paper stands out for its innovative use of 3D technology and psychographic data, and the second paper excels in its applicability to e-commerce optimization and handling of sparse data sets.
My models, though less advanced in terms of technology (ANN and GAN), cover a broad range of predictive capabilities, from basic logistic regression to the more complex random forest, offering robustness in handling diverse datasets. Also, my models are foundational and versatile for general classification tasks, the deep learning methods in the second paper provide a more tailored approach, particularly effective in dealing with sparse data and achieving personalization in recommendations. Both approaches have their unique strengths, with my models offering a broad applicability and the papers' methods showcasing advanced, specialized solutions for the fashion industry.

5 Results and Conclusions

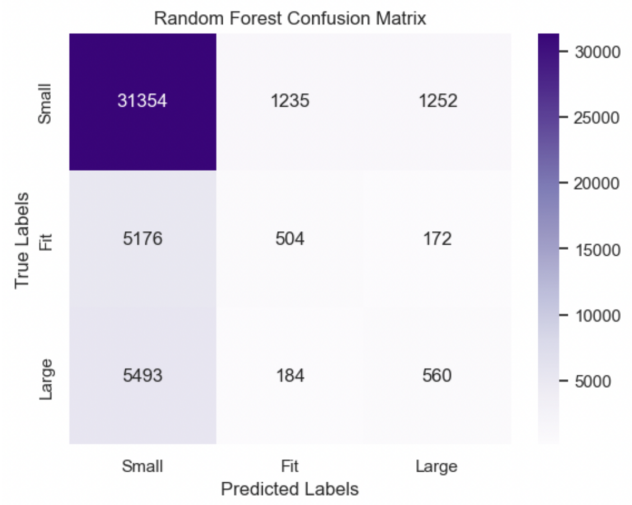
To perform validity assessment of all three models, I generated confusion matrices.The following is the confusion matrix for my first baseline model: Logistic Regression
This matrix shows that the model is particularly good at predicting the 'fit' class but has some challenges with the 'small' and 'large' classes. There are notable instances of misclassification between these two classes.



The following is the confusion matrix for my second baseline model: Naive Bayes Classifier Similar to the Logistic Regression model, the Naive Bayes classifier performs well in predicting the 'fit' class. However, it also struggles with the 'small' and 'large' classes, with a significant number of misclassifications.



The following is the confusion matrix for my Random Forest Classifier model:
The confusion matrix reveals that, like the previous models, Random Forest performs well in predicting the 'fit' class but struggles with the 'small' and 'large' classes. There is a notable number of instances where 'small' and 'large' classes are misclassified as 'fit', indicating potential challenges in



distinguishing between these classes.
Going back to table 7: *Performance Metrics of Random Forest Classifier with Different Paramet Settings* I used different values for ***n estimators*** and ***max depth*** for better model performances. My choices and what those parameter values mean are explained as follows:

Values for n estimators (Number of Trees):

- 10, 30: A low number of trees, very fast, but may not capture complex patterns well.
- 50: A moderate number, better for capturing patterns without being too slow.
- 70: Generally a good balance between performance and computational time.
- 80: Starts to increase computational time, but can improve model accuracy.
- 100, 120: Higher number for potentially better performance, but longer to train.

Same for the other column in table 7

Values for max depth (Maximum Depth of Each Tree):

- None: No maximum depth, trees grow until all leaves are pure or contain less than the minimum samples required to split a node. Can lead to overfitting.
- 2, 5: Shallow trees, faster training, but might underfit.
- 8: Deeper trees, capturing more complex patterns.
- 10: Provides more depth without being excessively deep for most datasets
- 11, 12: Deeper trees, which might improve accuracy but increase the risk of overfitting.

The parameter tuning helped increase the accuracy of the initial Random Forest Classifier model from 70.58% to 73.78%. The precision went from 62.20% to 67.85%. The recall went up from 70.58% to 73.78%. The F-1 score went down from 64.43% to 62.89%. The following table gives a final comparison of metrics reported by the three models used in my project.

In conclusion, this study successfully demonstrates the application of machine learning models to predict clothing fit with notable accuracy. Among the evaluated models, Random Forest, despite its computational intensity, showed promising results, particularly in terms of precision and F1-score. However, each model exhibited unique strengths and weaknesses, with Logistic Regression and Naive Bayes providing valuable baseline comparisons. Going forward, further improvements can be achieved through advanced feature engineering, exploring more sophisticated models like deep learning, and implementing techniques to handle class imbalance more effectively. Additionally, continuous refinement of the models with updated and more diverse datasets can enhance their predictive power. The integration of customer feedback loops and real-time data analytics could also provide dynamic adaptability to changing fashion trends and customer preferences, ultimately driving better customer satisfaction and business success.

Model	**Acc**	**Prec**	**Recall**	**F1**
Log. Regr.	73.56%	58.79%	73.57%	62.56%
Naive Bayes	73.57%	58.00%	73.57%	62.62%
Random Forest	70.58%	62.14%	70.58%	64.40%

Table 8: Performance Metrics of Various Models

References

1. Decomposing fit semantics for product size recommendation in metric spaces
   Rishabh Misra, Mengting Wan, Julian McAuley
   RecSys, 2018
   pdf
2. A novel approach in predicting virtual garment fitting sizes with psychographic characteristics and 3D body measurements using artificial neural network and visualizing fitted bodies using generative adversarial network
   Nga Yin Dik, Paul Wai Kei Tsang, Ah Pun Chan, Chris K.Y. Lo, and Wai Ching Chu
   report
3. A Deep Learning System for Predicting Size and Fit in Fashion E-Commerce
   Abdul-Saboor Sheikh, Romain Guigores, Evgenii Koriagin, Yuen King Ho, Reza Shirvany, Roland Vollgraf, Urs Bergmann
   report