AI-Driven Clothing Fit Prediction Using Machine Learning

Table of Contents

Introduction

In the evolving landscape of fashion and e-commerce, selecting the right clothing fit remains one of the most persistent challenges faced by consumers and retailers alike. With the proliferation of online shopping, customers are increasingly reliant on digital platforms to make purchasing decisions, yet a significant percentage of clothing returns stem from improper fit, leading to substantial financial losses for businesses and dissatisfaction among consumers. Traditional size charts and generic measurement guides often fail to accommodate the nuances of individual body shapes, fabric elasticity, and brand-specific

sizing variations, creating a fundamental gap between customer expectations and actual product fit. This inconsistency has propelled the need for more sophisticated, data-driven solutions to improve clothing recommendations and enhance customer experience. Currently, some retailers utilize user-generated reviews and body measurements to refine their recommendations, while others employ artificial intelligence (AI) and machine learning (ML) models to predict the best fit based on historical data. However, many existing solutions lack precision, personalization, and real-time adaptability, leading to continued inefficiencies. The Clothing Fit Predictor leverages advanced ML algorithms trained on diverse user attributes such as bust size, height, weight, body type, and rental reasons to enhance predictive accuracy. By incorporating large-scale user reviews and historical rental data, the model refines its learning to identify patterns and trends in fit preferences, significantly reducing return rates and enhancing customer satisfaction. This project aims to bridge the gap between consumers and retailers by providing an intelligent, automated, and highly accurate fit recommendation system, ultimately improving online shopping experiences and optimizing inventory management for businesses. Through the integration of machine learning models, natural language processing (NLP) for review analysis, and data visualization for insightful trends, the Clothing Fit Predictor stands as a transformative tool in the retail sector. As the fashion industry continues to embrace digital transformation, such AI-powered solutions will play a pivotal role in reshaping the future of online shopping, ensuring a seamless, efficient, and customer-centric approach to apparel purchases.

Exploratory Data Analysis (EDA)

Data Description

```
Data Shape: (192544, 15)

Data Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192544 entries, 0 to 192543
Data columns (total 15 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   fit             192544 non-null  object
 1   user_id         192544 non-null  int64
 2   bust size       174133 non-null  object
 3   item_id         192544 non-null  int64
 4   weight          162562 non-null  object
 5   rating          192462 non-null  float64
 6   rented for      192534 non-null  object
 7   review_text     192544 non-null  object
 8   body type       177907 non-null  object
 9   review_summary  192544 non-null  object
 10  category        192544 non-null  object
 11  height          191867 non-null  object
 12  size            192544 non-null  int64
 13  age             191584 non-null  float64
 14  review_date     192544 non-null  object
dtypes: float64(2), int64(3), object(10)
memory usage: 22.0+ MB
None
```

The dataset consists of 192,544 records and 15 features, capturing user interactions with rented clothing items. It includes numerical, categorical, and text-based attributes relevant to fit prediction. Key features include fit (target variable), bust size, weight, height, body type, rating, and rental purpose. Certain features like bust size, weight, body type, and height contain missing values, which need preprocessing. The dataset also includes textual fields such as review_text and review_summary, which can provide insights using Natural Language Processing (NLP). Overall, this dataset offers a comprehensive view of user preferences and fit experiences, making it suitable for building a predictive fit model.

Data Cleaning and Preprocessing

To ensure high-quality predictions, extensive data cleaning and preprocessing steps were applied. First, the weight column was converted to numerical format by stripping non-numeric characters, while height was transformed from feet-inches format to inches. Missing values in numerical columns such as weight, rating, height, and age were imputed using the median, while categorical features like bust size, rented for, and body type were filled with the mode. A new feature, BMI, was derived using height and weight. The bust size was further split into numeric and cup size components for better analysis. Date fields were converted to datetime format, extracting year and month for trend analysis. Finally, outlier detection was performed using Z-score filtering, and extreme values were removed

to improve model robustness. These steps ensured that the dataset was clean, well-structured, and optimized for machine learning models.

Descriptive Statistics & Insights

Statistic	User ID	Item ID	Rating	Size	Age	Weight (lbs)	Height (in)	BMI	Bust No.	Review Year	Review Month	Word Count
Count	180172	180172	180172	180172	180172	180172	180172	180172	180172	180172	180172	180172
Mean	499241.29	1.05M	9.21	11.39	33.38	135.30	65.30	22.33	34.08	2015.70	6.86	58.47
Min	9	123373	6.00	0.00	14.00	80.00	58.00	14.63	28.00	2010.00	1.00	1.00
25%	250351.5	197170	8.00	5.00	29.00	125.00	63.00	20.52	34.00	2015.00	4.00	28.00
50%	498650	958423	10.00	12.00	32.00	135.00	65.00	21.95	34.00	2016.00	7.00	50.00
75%	751284.8	1.69M	10.00	16.00	37.00	145.00	67.00	23.80	34.00	2017.00	10.00	79.00
Max	999997	2.97M	10.00	37.00	57.00	197.00	73.00	32.12	48.00	2018.00	12.00	398.00
Std Dev	289300.4	806751.2	1.19	7.08	7.15	16.69	2.63	2.62	1.55	1.33	3.38	42.90

The dataset consists of 180,172 records, providing insights into various user and review attributes. The average rating is 9.21, with a minimum of 6 and a maximum of 10, indicating mostly high ratings. The size variable has a mean of 11.39, but varies widely (0 to 37). The age distribution ranges from 14 to 57, with a median of 32, suggesting that most reviewers are young adults.

The BMI values range from 14.63 to 32.12, with a mean of 22.33, which is within a normal weight category. The bust size has a mean of 34.08, with most values around 34 (25th–75th percentile). The review word count varies significantly, from 1 to 398 words, indicating diverse review lengths.

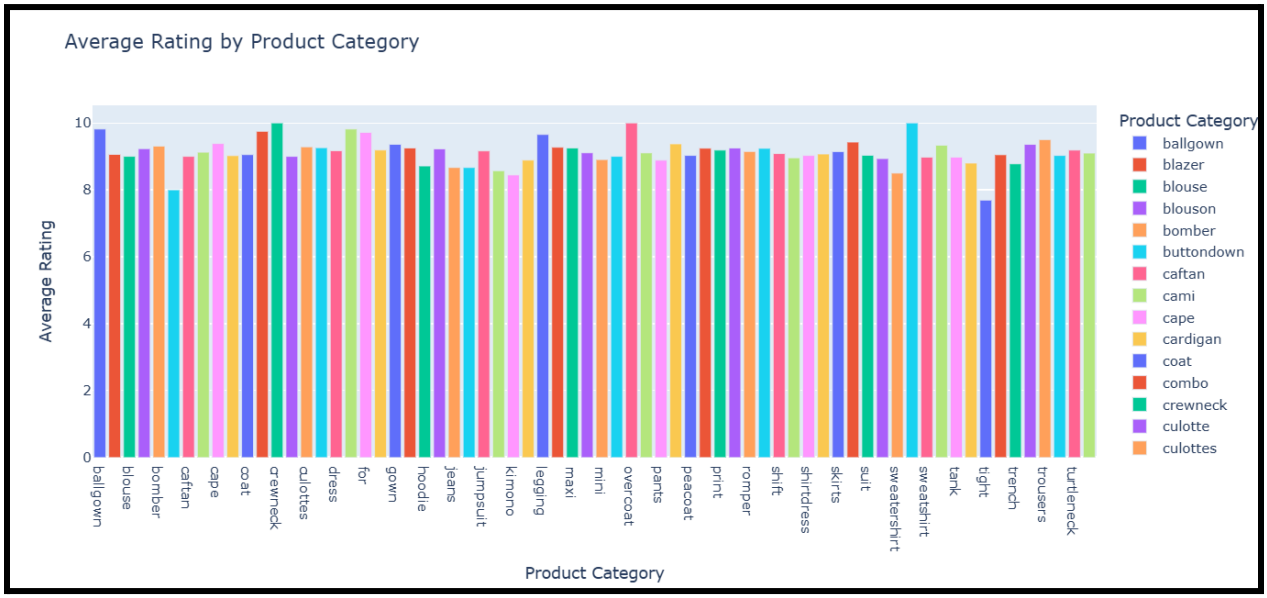
The data spans from 2010 to 2018, with the majority of reviews concentrated around 2015-2017. The standard deviation in most numeric features shows moderate variation, except for review_word_count, which has high variability.

Data Visualization

Word Cloud of Review Text

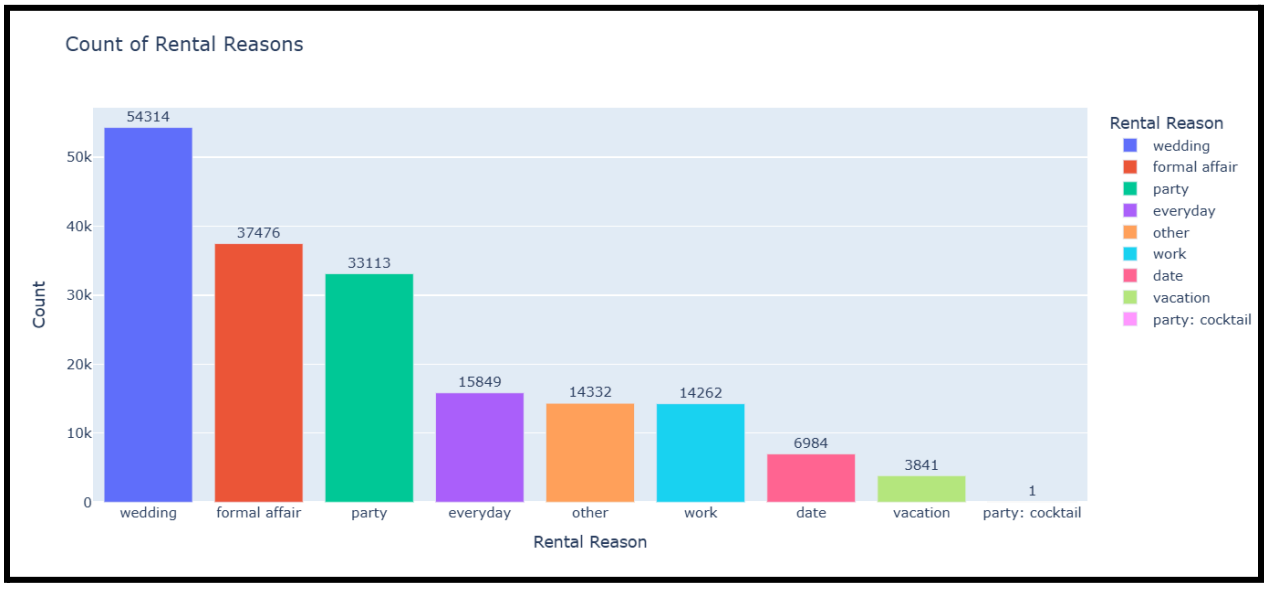
The word cloud of review texts highlights key themes and sentiments expressed by customers regarding their rented clothing. Prominent words such as "dress," "wore," "wear," "size," and "fit" indicate that fit and comfort are primary concerns for users. Words like "perfect," "comfortable," "flattering," and "compliment" suggest positive experiences when sizing and style expectations are met. Conversely, terms like "tight," "short," and "issue" may hint at occasional fit mismatches. The presence of "color," "fabric," and "material" underscores the importance of fabric quality in customer satisfaction. Businesses can leverage these insights to enhance product recommendations, optimize size charts, and refine marketing strategies for a more personalized shopping experience.

Average Rating by Product Category



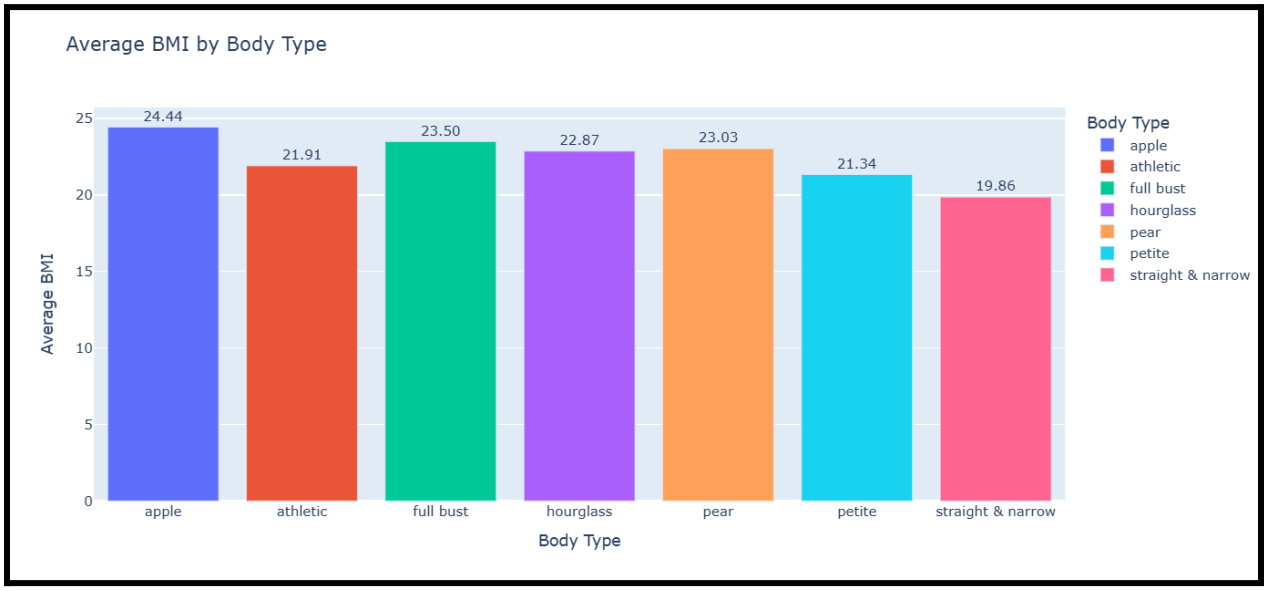
The bar chart illustrates the average rating across different product categories, showing customer preferences and satisfaction levels. Most categories maintain consistently high ratings (above 8), indicating overall positive experiences. Notably, ball gowns, jumpsuits, and certain coats receive the highest ratings, suggesting strong customer satisfaction. Some categories like sweatshirts and tight-fitting clothing show slightly lower ratings, possibly due to sizing or comfort issues. The diversity in ratings highlights varying expectations for different apparel types. Businesses can use these insights to enhance quality control, optimize inventory, and refine recommendations based on customer preferences for high-rated categories.

Rental Reasons Count



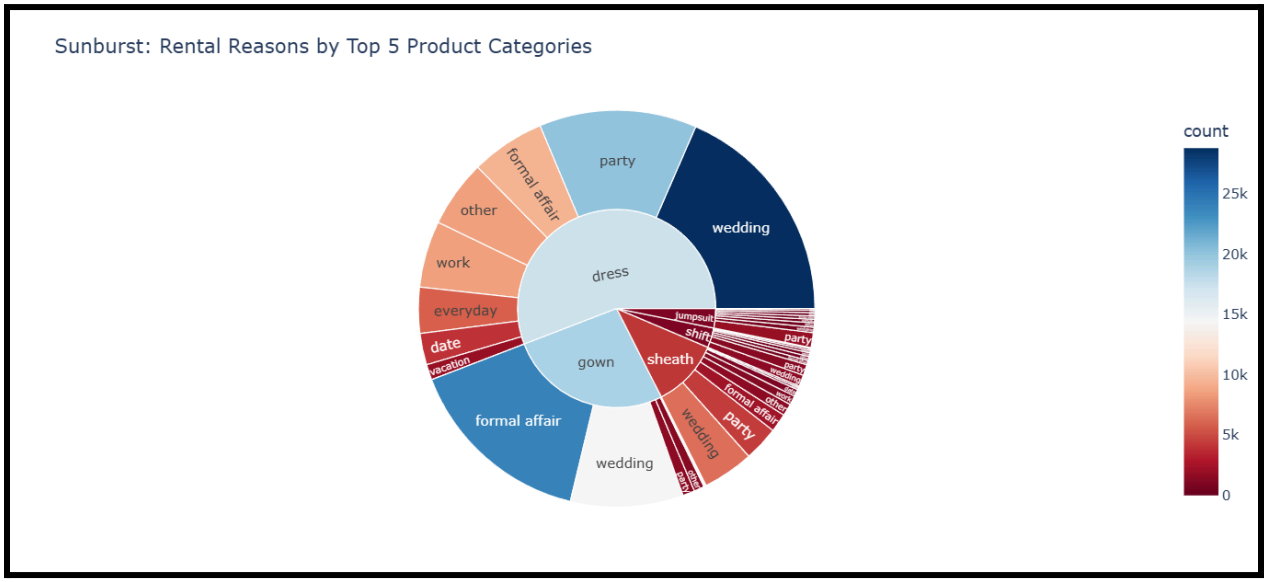
The bar chart displays the count of rental reasons, highlighting the most common purposes for clothing rentals. Weddings (54,314) dominate as the primary reason, followed by formal affairs (37,476) and parties (33,113), indicating that customers frequently rent high-end and event-specific outfits. Everyday wear (15,849) and work attire (14,262) show moderate demand, reflecting occasional rentals for professional settings. Vacation (3,841) and date night (6,984) have lower counts, suggesting that casual and short-term rentals are less popular. Businesses can leverage these insights to optimize inventory, prioritize high-demand categories, and tailor marketing strategies for key rental occasions.

Average BMI by Body Type



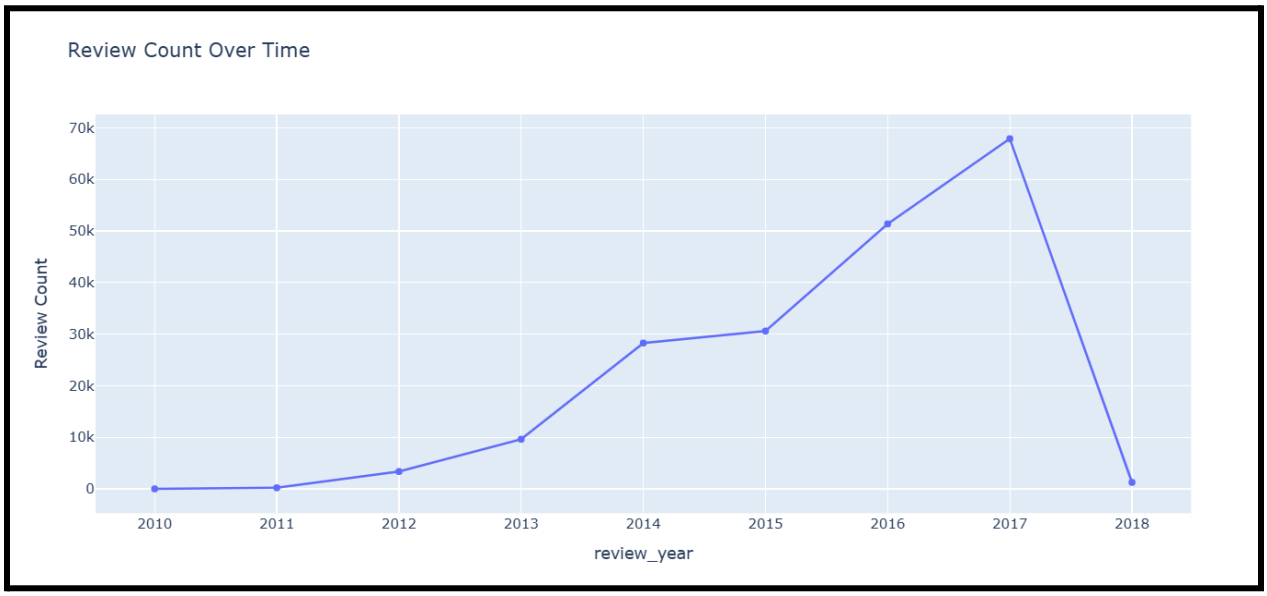
The bar chart presents the average BMI across different body types, highlighting variations in body composition among users. Apple-shaped individuals have the highest average BMI (24.44), indicating a tendency for weight distribution around the midsection. Full bust (23.50) and pear-shaped (23.03) body types also have relatively high BMIs. In contrast, straight & narrow (19.86) and petite (21.34) body types show lower average BMIs, suggesting a leaner frame. Understanding BMI distribution by body type helps businesses improve fit recommendations and size chart accuracy. These insights can optimize product sizing, ensuring better customer satisfaction and reduced return rates.

Rental Reasons by Top 5 Product Categories



The sunburst chart illustrates rental reasons across the top 5 product categories, providing insights into customer preferences. Dresses dominate, particularly for weddings and parties, reflecting high demand for formal and event-based attire. Gowns are primarily rented for formal affairs and weddings, reinforcing their association with luxury events. Sheath, jumpsuit, and shift dresses cater to diverse occasions, including work, everyday wear, and parties. The intensity of the colors represents rental volume, with weddings being the most common reason for rentals. Businesses can use these insights to stock high-demand styles and improve targeted marketing for occasion-based rentals.

Review Count over Time



The line chart illustrates the review count over time, showing a steady rise in user engagement from 2012 to 2017. A sharp increase is observed between 2014 and 2017,

peaking at nearly 70,000 reviews in 2017, indicating a growing user base and higher interaction levels. However, there is a significant decline in 2018, likely due to a decrease in rentals, platform changes, or missing data. The overall trend suggests increasing customer reliance on reviews for decision-making. Businesses can leverage these insights to enhance customer feedback strategies and improve engagement through targeted marketing efforts.

Feature Engineering

Feature Selection

To improve model efficiency and eliminate irrelevant data, certain columns were removed before model training. User-specific identifiers such as user_id and item_id were dropped as they do not contribute to predicting clothing fit. Text-based fields, including review_text and review_summary, were removed due to their unstructured nature. Date-related columns like review_date, review_year, and review_month were excluded to prevent temporal biases. Additionally, redundant features such as height and weight were omitted after deriving more informative metrics like BMI. The final set of features includes only the most relevant attributes to enhance model accuracy and efficiency in predicting clothing fit.

Feature Encoding and normalization

To prepare the data for machine learning models, categorical variables were encoded using Label Encoding. Features such as bust size, body type, cup size, rented for, and category were transformed into numerical values, allowing the model to interpret categorical distinctions effectively. The target variable (fit) was also label-encoded to ensure compatibility with classification algorithms. After encoding, numerical features were standardized using StandardScaler, which normalizes them to a common scale, improving model stability and performance. This preprocessing step ensures that all features contribute equally to model predictions, avoiding bias from disproportionately scaled values.

Handling Class Imbalance

The dataset exhibited class imbalance, where the majority class significantly outweighed the minority classes, potentially leading to biased model predictions. To address this, Synthetic Minority Over-sampling Technique (SMOTE) was applied, which generates synthetic samples for the underrepresented classes to create a more balanced distribution. This technique enhances model generalization by preventing it from being biased toward the majority class. By resampling the dataset using SMOTE, the model is trained on a more diverse set of instances, leading to improved classification performance and better fit predictions for all user groups.

Model Development

Model Selection

The Random Forest classifier was selected for predicting clothing fit due to its robustness, accuracy, and ability to handle both numerical and categorical data. Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions, reducing the risk of overfitting while improving generalization. It is highly effective in handling missing values, outliers, and complex feature interactions, making it ideal for our dataset, which contains various categorical and numerical variables. Additionally, Random Forest provides feature importance scores, helping to identify key attributes influencing fit predictions. Given its efficiency in handling large datasets and its ability to deliver high accuracy, it was the most suitable choice for this classification task.

Training the Model

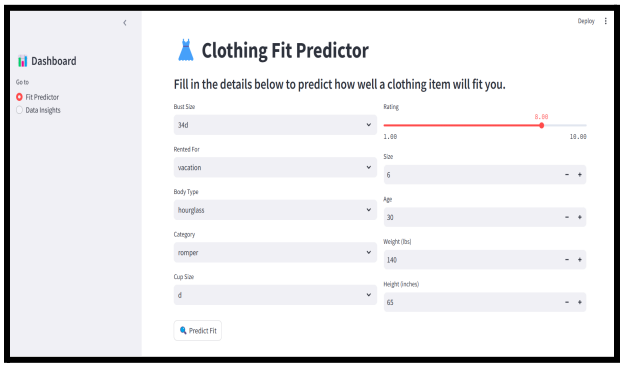
The Random Forest classifier was trained using 200 estimators, with a maximum depth of 20 and a minimum sample split of 5, ensuring a balance between performance and generalization. The random_state parameter was set to 42 for reproducibility, and n_jobs=-1 was used to leverage parallel processing, optimizing computational efficiency. The model was trained using the preprocessed dataset, including balanced classes through SMOTE, standardized numerical features, and label-encoded categorical variables. By learning from diverse patterns in user attributes, the model enhances its ability to predict clothing fit accurately. This setup ensures a well-generalized classifier capable of delivering high accuracy while maintaining robustness across different customer profiles.

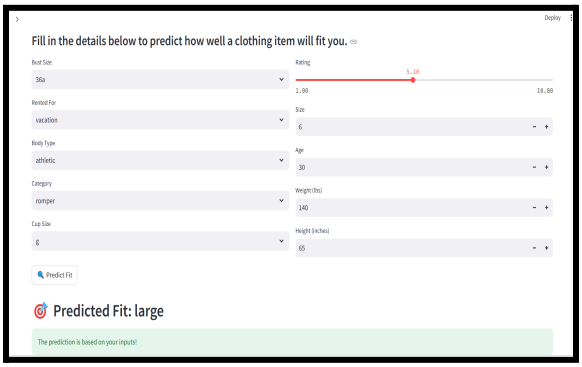
Model Performance Metrics

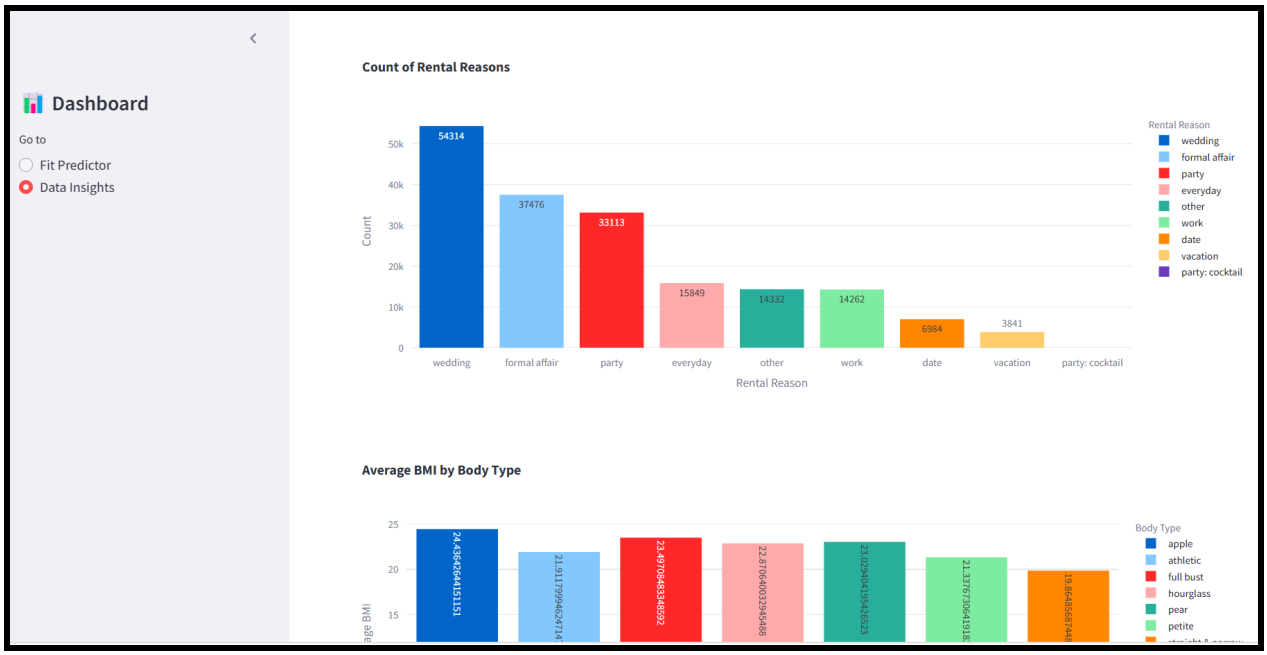
```
Accuracy: 0.78063620704611108
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.83      0.79     26899
           1       0.82      0.73      0.77     26899
           2       0.78      0.78      0.78     26899

    accuracy                           0.78     80697
   macro avg       0.78      0.78      0.78     80697
weighted avg       0.78      0.78      0.78     80697

Confusion Matrix:
 [[22394  1932  2573]
 [ 3815 19745  3339]
 [ 3524  2519 20856]]
```

demonstrating strong predictive capability. The classification report shows a balanced precision, recall, and F1-score, with values around 0.78 for all classes, indicating reliable performance across different fit categories. The confusion matrix highlights that the model correctly classifies most instances while maintaining a good balance in predicting all classes. The weighted average F1-score of 0.78 confirms the model's robustness in handling multi-class classification. The results suggest that the model effectively generalizes fit predictions while maintaining fairness across different user profiles, making it a practical tool for real-world applications.

Model Deployment on Streamlit







The Clothing Fit Predictor is deployed using Streamlit, an interactive web-based application framework that allows users to input their attributes and receive a predicted fit in real-time. The deployment involved loading the trained Random Forest model, along with preprocessing objects such as scalers and label encoders, to ensure consistent data transformation during predictions. The application features an intuitive UI with dropdowns for categorical features (e.g., bust size, body type, rental reason) and sliders for numerical attributes (e.g., weight, height, rating).

The app is structured into two sections: Fit Predictor and Data Insights. The Fit Predictor section allows users to input their measurements and receive an immediate prediction of fit. The Data Insights dashboard offers visualizations, including bar charts, correlation heatmaps, and word clouds, providing users with analytics on rental trends, body types, and review patterns. By leveraging Google Drive for data retrieval, the application ensures scalability and remote data accessibility. This deployment allows for seamless interaction, enabling retailers and customers to make more informed purchasing decisions.

Conclusion

The AI-Driven Clothing Fit Prediction system leverages machine learning to address one of the most persistent challenges in e-commerce—accurate clothing fit recommendations. By analyzing diverse user attributes, such as height, weight, body type, and historical rental data, the model significantly enhances fit predictions, reducing return rates and improving customer satisfaction. The integration of NLP for review analysis and data visualization further refines insights, allowing businesses to optimize inventory and marketing strategies. The deployment of the model using Streamlit ensures accessibility and real-time usability, making it a practical solution for retailers and consumers. With an F1-score of 0.78, the Random Forest-based model demonstrates strong predictive accuracy, effectively

balancing precision and recall. As AI-driven solutions continue to reshape online shopping, this project highlights the potential of machine learning in delivering personalized, data-driven clothing recommendations, ultimately transforming the retail experience and bridging the gap between customer expectations and product fit.