**Customer Clothing Fit Analysis & Predictive Modeling**

**Overview**

This project analyzes customer clothing fit data from Rent The Runway to improve clothing recommendations and fit predictions. Using data preprocessing, exploratory data analysis (EDA), feature engineering, and machine learning, we create a model that predicts fit preferences based on user characteristics.

**Tools & Technologies Used**

* **Programming Language**: Python
* **Libraries**: NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, NLTK
* **Machine Learning Algorithms**: Logistic Regression, KNN, Naïve Bayes, Random Forest, Stacking Classifier, AdaBoost, Gradient Boosting, Decision Tree
* **Dimensionality Reduction**: Principal Component Analysis (PCA)

**Project Breakdown**

**1. Data Collection & Loading**

* The dataset (renttherunway.csv) was loaded using Pandas.
* Initial exploration included checking dataset shape, column types, and summary statistics.

**2. Data Cleaning & Preprocessing**

* Removed redundant columns: user\_id, item\_id, review\_date, and Unnamed: 0.
* **Handling Missing Values**:
  + Rows with more than two missing values were dropped.
  + Missing values in categorical columns (e.g., bust size, body type) were imputed using the most frequent values.
  + age and heightINCH missing values were replaced with their respective mean values.
  + weight was predicted using an AdaBoost model since it had a high percentage of missing values.
* **Feature Engineering**:
  + Converted height from feet-inches format to total inches.
  + Transformed bust size into numerical values for better analysis.
  + Standardized numerical features using StandardScaler.

**3. Exploratory Data Analysis (EDA)**

* Identified duplicate values (none found).
* Analyzed unique values in categorical features (rented for, body type, etc.).
* Checked for outliers in age using box plots and replaced extreme values using the interquartile range (IQR) method.
* Visualized distribution of key variables using histograms.

**4. Sentiment Analysis on Reviews**

* Used NLTK SentimentIntensityAnalyzer to analyze customer reviews.
* Created a new feature ratio, which quantified the sentiment impact based on the number of words.

**5. Encoding Categorical Variables**

* One-hot encoding was applied to categorical columns: fit, rented for, body type, and category.

**6. Machine Learning Model to Predict Weight**

* **Features Used**: size, heightINCH, bust size new
* **Target**: weight
* **Model Used**: AdaBoost Regressor
* The model achieved **85% accuracy** within a 5% margin of error.
* Missing weight values were replaced with model-predicted values.

**7. Dimensionality Reduction with PCA**

* Performed PCA to reduce the feature set while maintaining **90% variance**.
* Identified key components contributing the most to variance.

**8. Model Building & Evaluation**

* Implemented multiple classification models (Logistic Regression, Random Forest, KNN, Naïve Bayes, Stacking Classifier).
* Compared performance based on **accuracy, confusion matrix, and feature importance**.

**Results & Insights**

* Successfully cleaned and transformed raw data into a structured dataset.
* Created a predictive model for **customer weight estimation**, handling missing values effectively.
* Analyzed customer sentiments using text data.
* Applied PCA for **dimensionality reduction** to improve model efficiency.
* Built and tested multiple classification models for **clothing fit analysis**.

**Conclusion**

This project demonstrates **data wrangling, feature engineering, machine learning, and sentiment analysis** to enhance clothing fit predictions. By leveraging **advanced modeling techniques**, we optimized predictions and improved customer recommendations.

Github 🡪 https://github.com/yashzob/dataScienceRentTheDress