Predictive Modeling for User Satisfaction in Rental Clothing Recommendations

CSE 258R - Assignment 2
University of California San Diego

***Abstract*—This study explores various predictive models for user ratings in a clothing rental recommendation system. We analyze a dataset of 192,544 reviews, implementing baseline models, linear regression approaches, and advanced techniques like TF-IDF, SVD, and custom latent factor models. Our findings reveal that linguistic features, particularly exclamation marks in reviews, significantly outperform simple predictors. The custom latent factor model achieves the best performance with an MSE of 0.317, demonstrating the effectiveness of tailored approaches in handling data sparsity and rating normalization. The resulting recommender system can generate personalized top-5 item recommendations with estimated ratings for individual users, showcasing its potential for enhancing user experience in e-commerce and rental platforms.**

***Index Terms*—Recommender Systems, Web Mining, Machine Learning, Prediction Models**

Fig. 1: AI-generated image showing the widespread adoption of recommender systems in shopping.

I. Introduction

Recommender systems play a crucial role in predicting and suggesting clothing products that a human would like, addressing several key challenges in the online fashion industry:

A. Personalization in a Vast Product Space

The online clothing market offers an overwhelming array of choices, making it difficult for consumers to find items that match their preferences. Recommender systems help tackle this problem by:

- Analyzing individual user behavior and preferences
- Identifying patterns across similar users
- Suggesting relevant items tailored to each person's taste

This personalization is essential because clothing preferences are highly subjective and vary greatly from one individual to another.

B. Overcoming Information Overload

With thousands of products available, users can easily become overwhelmed when browsing online clothing stores. Recommender systems alleviate this issue by:

- Filtering out irrelevant options
- Highlighting items likely to appeal to the user
- Reducing the cognitive load of decision-making

This curated approach helps users discover products they might otherwise miss, improving their shopping experience and satisfaction.

C. Addressing Fit and Style Challenges

One of the biggest hurdles in online clothing retail is predicting whether an item will fit and suit a particular customer. Recommender systems can help by:

- Analyzing past purchase and return data
- Considering user-provided measurements and preferences
- Factoring in style compatibility and current fashion trends

By incorporating these elements, recommenders can suggest items more likely to fit well and align with the user's personal style.

D. Improving Customer Experience and Retention

Effective clothing recommendations can significantly enhance the overall shopping experience by:

- Saving users time in searching for suitable items

- Introducing them to new styles or brands they might enjoy
- Increasing the likelihood of successful purchases

This improved experience can lead to higher customer satisfaction, increased loyalty, and reduced return rates.

E. Driving Sales and Business Growth

From a business perspective, recommender systems are invaluable for:

- Increasing conversion rates
- Encouraging higher average order values through complementary suggestions
- Helping to clear inventory by promoting relevant but less visible items

These benefits make recommender systems a critical tool for online fashion retailers looking to boost their bottom line and compete in a crowded marketplace like that seen in Fig. 1. In conclusion, recommender systems are essential in the online clothing industry, helping to bridge the gap between the vast array of available products and each individual's unique preferences and needs. By leveraging data and advanced algorithms, these systems enhance the shopping experience, improve customer satisfaction, and drive business success in an increasingly competitive digital landscape.

II. Literature Survey

Recommendation systems have become increasingly important in the fashion industry, particularly with the rise of online shopping and clothing rental services [1]. This section reviews relevant literature in the field of fashion recommendation systems, focusing on collaborative filtering, content-based approaches, and evaluation techniques.

A. Collaborative Filtering in Fashion Recommendation

Collaborative filtering has been widely used in fashion recommendation systems. [2] examined various collaborative filtering algorithms for clothing recommendations in e-commerce, highlighting the importance of accounting for user biases in rating behaviors. [3] proposed a personalized outfit generation system for fashion recommendation at Alibaba iFashion, demonstrating the effectiveness of collaborative filtering techniques in large-scale e-commerce platforms.

B. Content-Based Approaches

Content-based filtering approaches have also been explored in fashion recommendation systems. A study [4] proposed an enhanced content-based fashion recommendation system using deep ensemble classifiers. This approach combines multiple pre-trained models to improve classification accuracy and leverages transfer learning techniques for better performance on benchmark datasets. [5] investigated the use of neighbor-constrained embedding learning to understand fashion trends from street photos, showcasing the potential of content-based approaches in capturing visual features and style preferences.

C. Hybrid Models

Hybrid models that combine collaborative filtering and content-based approaches have shown promise in addressing the limitations of individual techniques. [6] proposed an LSTM-based dynamic customer model for fashion recommendation, integrating temporal dynamics with collaborative filtering.

D. Evaluation Techniques

Evaluating the effectiveness of recommendation systems is crucial for their improvement. A comprehensive survey [7] introduces five key types of metrics for holistic evaluation: similarity metrics, candidate generation metrics, predictive metrics, ranking metrics, and business metrics. This framework provides a nuanced approach to assessing recommendation systems, considering both technical performance and business objectives.

E. Challenges in Fashion Recommendation

Fashion recommendation systems face unique challenges, including data sparsity and the cold-start problem. Previous research [8] highlights these issues and proposes novel approaches to address them in the context of fashion rental services. The study emphasizes the importance of considering individual item attributes and rental histories in developing effective recommendation systems. In conclusion, the field of fashion recommendation systems is rapidly evolving, with researchers exploring various techniques to improve recommendation accuracy, address domain-specific challenges, and enhance user satisfaction. Future research directions may include developing more sophisticated hybrid models that combine collaborative filtering, content-based approaches, and advanced machine learning techniques to better capture the complexities of user preferences in fashion.

III. Overview of Dataset

A. Summary of Dataset

The dataset [1] contains 192,544 entries of user reviews for rented clothing items. Each entry includes detailed information about the user, the rented item, and the user's experience with it. Key aspects of the dataset include:

User Information:

- **Demographics:** Age, height, weight, body type, and bust size.
- **Body Types:** Self-identified categories such as hourglass, athletic, pear, petite, full bust, straight & narrow, and apple.

Item Information:

- **Categories:** Wide range including dresses, gowns, rompers, jumpsuits, tops, sweaters, jackets, and leggings.
- **Sizes:** Range from 0 to 28, with some items using letter sizing (S, M, L, etc.).
- **Item IDs:** Unique identifier for each item.

Rental Information:

- **Rental Occasion:** Specified events such as weddings, vacations, formal affairs, parties, work events, and everyday wear.
- **Rental Date:** Inferred from the review date.

User Feedback:

- **Ratings:** Scale from 1 to 10.
- **Fit:** Indications of whether the item fit as expected, was small, or was large.
- **Review Text:** Detailed written feedback about fit, quality, comfort, and overall experience.
- **Review Summary:** Brief title or summary of the review.

Data Format and Structure:

- **Format:** JSON Lines (each line is a JSON object).
- **Example Fields:** fit, user_id, bust size, item_id, weight, rating, rented for, review_text, body type, review_summary, category, height, size, age, review_date.

Key Features:

- Detailed user feedback through rich, qualitative data in reviews.
- Diverse range of clothing items for cross-category analysis.
- Contextual information on rental occasions.
- Comprehensive size and fit data across different body types.
- Temporal aspect enabling trend analysis over time.

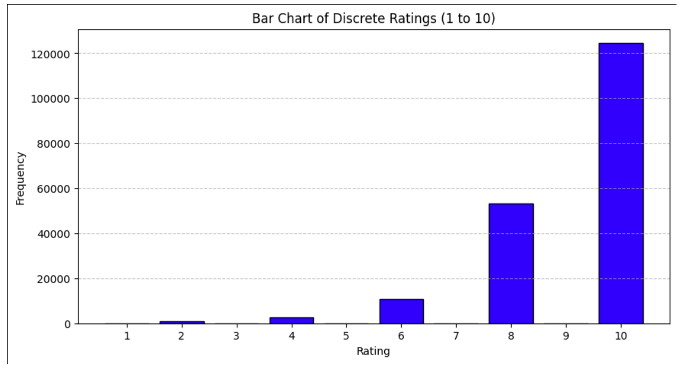

Fig. 2: Histogram showing the distribution of ratings for the original dataset

This dataset is particularly valuable for analyzing consumer preferences in the clothing rental market, understanding fit issues across different body types, and developing recommendation systems for fashion items. It also provides insights into the relationship between user characteristics, item attributes, and overall satisfaction with rented clothing.

B. Statistical Summary

Metric	Value
Minimum rating	2
Mean rating	9.09
Mode rating	10
Median rating	10
Standard deviation	1.43

TABLE I: Pre-Adjustment Rating Statistics

1) Pre-Adjustment Statistics: The pre-adjustment statistics reveal a highly skewed distribution of ratings:

- **Range:** With a minimum of 2 and a maximum of 10 (implied by the mode), the ratings span 8 points on the scale.
- **Central Tendency:** The mean (9.09), median (10), and mode (10) are all clustered near the maximum possible rating, indicating an extremely positive bias.
- **Dispersion:** The standard deviation of 1.43 suggests moderate variability, but given the high mean, this indicates that most ratings are concentrated at the upper end of the scale.
- **Skewness:** The fact that the mean (9.09) is lower than both the median and mode (10) indicates a left-skewed distribution, with a long tail towards lower ratings.

This distribution suggests a strong positive bias in user ratings, possibly due to factors such as selection bias (satisfied customers being more likely to leave reviews) or a ceiling effect in the rating scale.

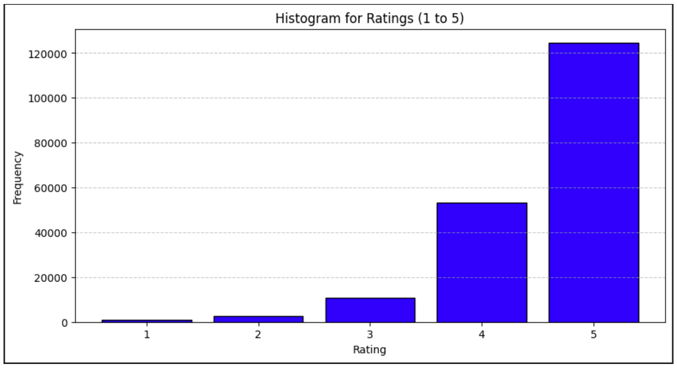

Fig. 3: Histogram showing the rating distribution after scaling

Metric	Value
Minimum rating	1
Mean rating	4.54
Mode rating	5
Median rating	5
Standard deviation	0.72

TABLE II: Post-Adjustment Rating Statistics

2) Post-Adjustment Statistics: After adjustment, likely to a 5-point scale, the distribution shows more balance:

- **Range:** The new scale spans from 1 to 5, providing a full range of rating options.
- **Central Tendency:** The mean (4.54) is now closer to the center of the scale, while the median and mode (both 5) still indicate a positive bias, but less extreme than before.
- **Dispersion:** The standard deviation has decreased to 0.72, suggesting a tighter clustering of ratings around the mean.
- **Skewness:** The relationship between mean, median, and mode suggests a slight left skew, but much less pronounced than in the pre-adjustment data.

The adjusted scale provides a more nuanced view of user satisfaction:

- The mean of 4.54 on a 5-point scale translates to approximately 90.8% satisfaction, closely mirroring the pre-adjustment mean of 9.09 on a 10-point scale.
- The reduced standard deviation (0.72) relative to the scale range indicates more consistent ratings, potentially allowing for finer differentiation between items or experiences.
- The persistence of a high median and mode (5) suggests that while the adjustment provides more granularity, the overall user sentiment remains highly positive.

This adjustment likely improves the utility of the ratings for analysis and comparison purposes, while maintaining the overall positive sentiment expressed by users in the original data.

IV. Analysis of Dataset

A. Correlation Analysis

Correlation studies were conducted to identify parameters that significantly influence user ratings. The results provide insights into the relationships between various factors and user satisfaction.

Parameters	Correlation with Rating
Age	-0.03519
Height	0.00174
Weight	-0.02068
Fit	0.24502
Purpose	0.01292
Review text (! Count)	0.18009

TABLE III: Correlation Coefficients with User Ratings

1) Interpretation of Results:

a) Weak or Negligible Correlations:

- **Age** (-0.03519): The slightly negative correlation suggests a very weak tendency for older users to rate items marginally lower, but the effect is negligible.
- **Height** (0.00174): With a correlation near zero, height appears to have virtually no linear relationship with ratings (Fig. 4).
- **Weight** (-0.02068): Similar to age, weight shows a very weak negative correlation, indicating a minimal tendency for heavier users to rate slightly lower.
- **Purpose** (0.01292): The purpose of rental (Fig. 5) has a negligible positive correlation, suggesting that the occasion for which an item is rented barely influences the rating.

These weak correlations indicate that demographic factors and rental purpose are not reliable predictors of user satisfaction in a linear model.

b) Moderate Correlations:

- **Fit** (0.24502): The strongest correlation observed, indicating that better-fitting items tend to receive higher ratings. This moderate positive correlation suggests that fit is a key factor in user satisfaction.
- **Review Text (! Count)** (0.18009): The moderate positive correlation between exclamation marks in

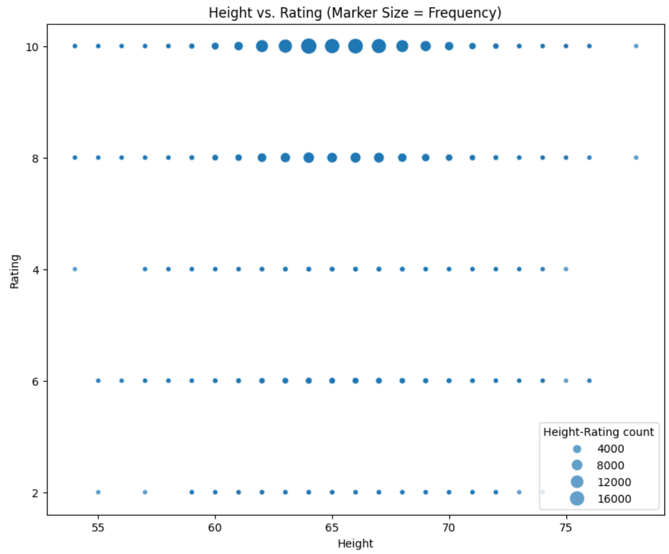

Fig. 4: Scatter plot denoting the correlation between height of an individual and the rating

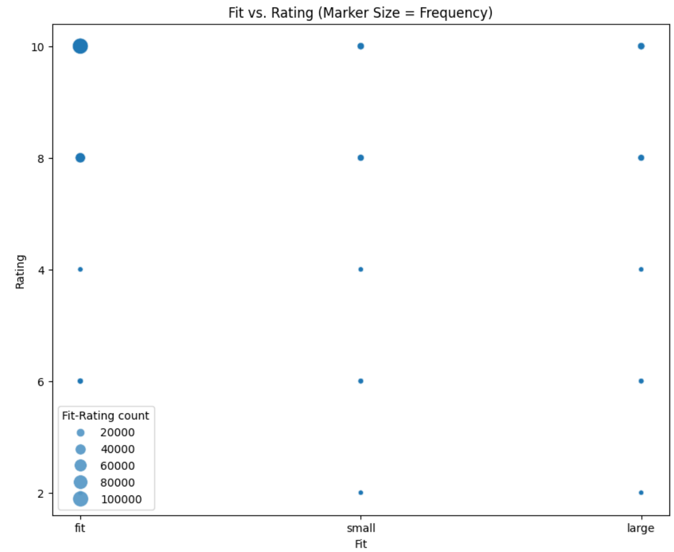

Fig. 6: Scatter plot denoting the correlation between the fit of an item and the rating

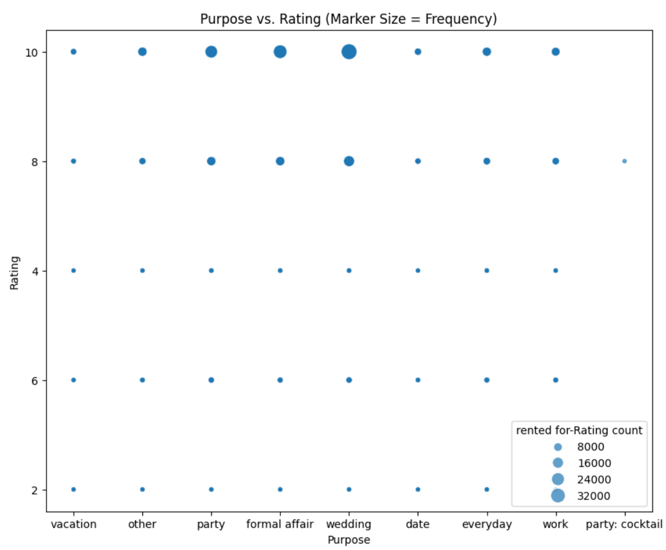

Fig. 5: Scatter plot denoting the correlation between the purpose of an item and the rating

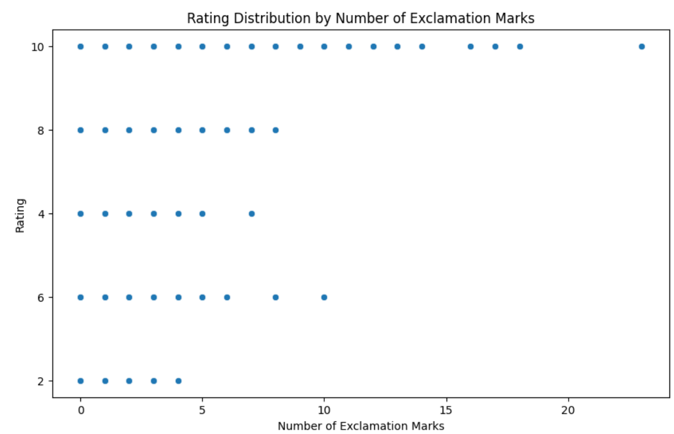

Fig. 7: Scatter plot denoting the correlation between the number of exclamation points in review and the rating

reviews and ratings suggests that users who express more enthusiasm in their reviews tend to give higher ratings.

2) Implications for Linear Regression Modeling: Based on these correlations, the most promising parameters for a linear regression model to predict ratings are:

1) **Fit** (0.24502): As the strongest predictor, fit should be a primary feature in the model (Fig. 6). Its moderate correlation suggests that improvements in fit could lead to notable increases in user satisfaction.
2) **Review Text (! Count)** (0.18009): While not as strong as fit, the correlation with exclamation marks in reviews could provide additional predictive power (Fig. 7). This feature might serve as a proxy for user enthusiasm or overall positive experience.

3) Further Considerations:

- **Non-linear Relationships**: The weak correlations for demographic factors don't necessarily mean these variables are irrelevant. They may have non-linear relationships with ratings that aren't captured by simple correlation coefficients.
- **Interaction Effects**: While individual correlations are weak for some variables, there might be signif-

icant interaction effects (e.g., between age and fit) that could be explored in more complex models.
- **Categorical Variables**: For categorical variables like 'Purpose', a more detailed analysis (e.g., ANOVA) might reveal differences in ratings across categories that aren't apparent in a simple correlation.
- **Review Text Analysis**: The correlation with exclamation marks suggests that more comprehensive text analysis (sentiment analysis, keyword extraction) could yield valuable predictive features.

This correlation analysis provides a foundation for feature selection in predictive modeling and highlights the importance of fit and user enthusiasm in determining satisfaction with rented clothing items. However, it also underscores the complexity of user satisfaction, suggesting that a combination of features and possibly non-linear modeling approaches may be necessary to fully capture the factors influencing user ratings.

B. Principal Component Analysis (PCA)

Principal Component Analysis was performed to understand the underlying structure of the dataset. The results provide insights into the relationships between various features and their contributions to the overall variance in the data.

Principal Component	Explained Variance Ratio
PC1	0.2086
PC2	0.1302
PC3	0.1182
PC4	0.1125
PC5	0.0967
PC6	0.0817
PC7	0.0809
PC8	0.0763
PC9	0.0698
PC10	0.0251

TABLE IV: Explained Variance Ratio by Principal Components

1) Explained Variance Ratio: Analysis of the explained variance ratio reveals (Fig. 8):

- The first principal component (PC1) accounts for 20.86% of the total variance, indicating that no single component dominates the dataset's variability.
- The top three components (PC1, PC2, PC3) collectively explain 45.7% of the variance, suggesting a complex data structure with multiple influential factors.

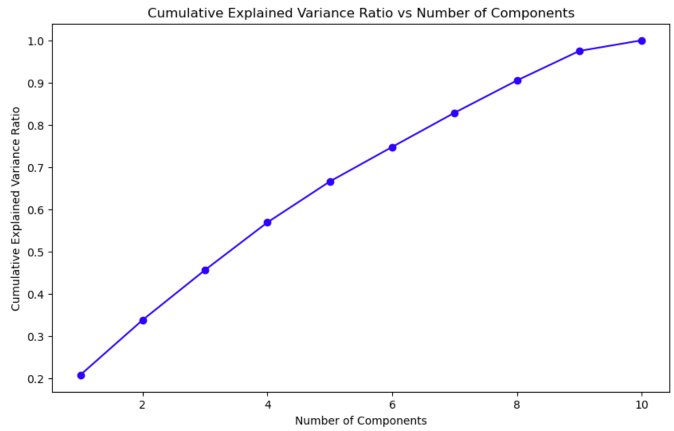

Fig. 8: Graph highlighting the explained variance ration for different features

- There is a gradual decrease in explained variance from PC1 to PC9, followed by a more significant drop for PC10.
- The first 9 components cumulatively account for 97.49% of the total variance, indicating that these components capture the majority of the dataset's information.

2) Feature Importance: Analysis of feature contributions to principal components reveals significant insights (Fig. 9):

- **PC1:** Strongly influenced by weight (0.612372) and size (0.592902), suggesting a primary axis of variation related to physical dimensions.
- **PC2:** Heavily impacted by fit (-0.526497) and rating (-0.550611), indicating a strong relationship between how well an item fits and user satisfaction.
- **PC3:** Shows substantial contributions from body type (0.580689) and bust size (0.494893), highlighting the importance of body shape in the data structure.
- **Other PCs:** While not explicitly quantified, subsequent components likely capture more nuanced relationships between features.

3) Key Insights: The PCA results reveal complex interrelationships within the dataset:

- **Physical Attributes:** Weight, size, and height are primary contributors to variance, particularly in PC1. This suggests that these physical characteristics are fundamental in explaining differences in user experiences and preferences.
- **Fit and Satisfaction:** The strong influence of fit and rating on PC2 underscores the critical relationship between how well a garment fits and customer

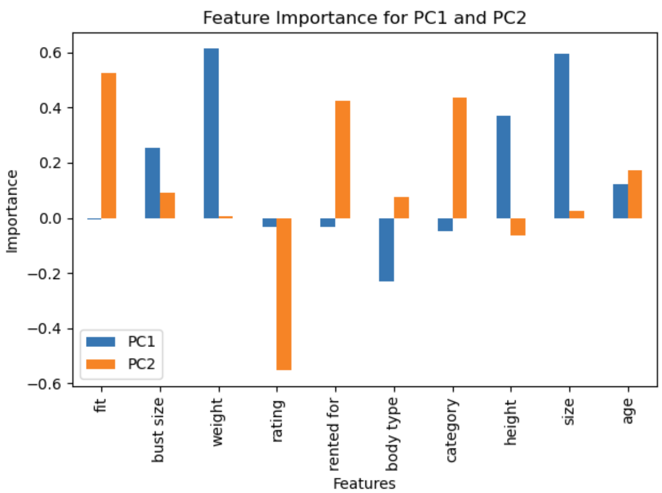

Fig. 9: Bar plot explaining the importance of different features for PC1 and PC2

satisfaction. This could be valuable for improving sizing recommendations and overall service quality.
- **Body Shape Considerations:** The prominence of body type and bust size in PC3 indicates that these factors play a significant role in differentiating user experiences, potentially affecting fit and style preferences.
- **Multifaceted User Experience:** The distribution of variance across multiple components suggests that the user experience in clothing rental is complex and influenced by a variety of factors. No single aspect dominates, highlighting the need for a holistic approach in analyzing and improving the service.
- **Review Characteristics:** While not explicitly quantified, the mention of review-specific attributes (e.g., exclamation marks) having moderate impact across components suggests that the way users express their opinions also contributes to the overall data structure.

These PCA results provide a foundation for further analysis, potentially guiding feature selection for predictive modeling, informing personalization strategies, and identifying key areas for service improvement in the clothing rental business.

V. Generating Baselines

To establish a foundation for predicting user ratings, we developed and analyzed several linear regression models. Each model uses different features to predict ratings, providing insights into the relative importance of various factors.

A. Fixed Value (Median) Model

This simplest model predicts a constant rating for all instances:

- Constant (c) = 5
- Mean Squared Error (MSE) = 0.51638951

The fixed value model serves as our baseline. Its relatively high MSE (0.51638951) indicates poor predictive power, which is expected given the variability in user ratings. This model's performance suggests that any model with an MSE below 0.51638951 offers some predictive value.

B. Length of Review Model

This model uses the length of the review text as a predictor:

- Slope (m) = -4.4853704566801365e-05
- Intercept (c) = 4.558203420946161
- MSE = 0.51629136

The slightly negative slope (-4.4853704566801365e-05) suggests a minimal inverse relationship between review length and rating. However, the MSE (0.51629136) is only marginally better than the fixed value model, indicating that review length alone is a weak predictor of user satisfaction.

C. Count of Exclamation Marks Models

1) Full Review Exclamation Count Model: This model uses the number of exclamation marks in the entire review:

- Slope (m) = 0.10068846
- Intercept (c) = 4.4551229906
- MSE = 0.49989404

The positive slope (0.10068846) indicates that more exclamation marks correlate with higher ratings. The lower MSE (0.49989404) suggests this model outperforms both the fixed value and review length models.

2) Summary Exclamation Count Model: This model focuses on exclamation marks in the review summary:

- Slope (m) = 0.1835002556
- Intercept (c) = 4.43549445
- MSE = 0.49644174

With a steeper positive slope (0.1835002556) and lower MSE (0.49644174), this model outperforms the full review exclamation count model, suggesting that exclamation marks in summaries are stronger indicators of user satisfaction.

D. Combined Exclamations and Review Length Model

This model incorporates both exclamation count and review length:

- Exclamation Count Coefficient (m1) = 0.114873339
- Review Length Coefficient (m2) = -0.00025662
- Intercept (c) = 4.5199050
- MSE = 0.49700863

This model achieves the lowest MSE (0.49700863) among all tested models. The positive coefficient for exclamation count (0.114873339) and slightly negative coefficient for review length (-0.00025662) suggest that enthusiastic, concise reviews tend to correlate with higher ratings.

E. Physical Characteristics Model

This model uses height, weight, and age to predict ratings:

- Height Coefficient (m1) = 0.00075905
- Weight Coefficient (m2) = -9.555753e-05
- Age Coefficient (m3) = -0.002927279
- Intercept (c) = 4.605015781
- MSE = 0.515731479553

While this model outperforms the fixed value model ($MSE = 0.515731479553 < 0.51638951$), it falls short of the linguistic feature models. The coefficients suggest slight positive correlation with height (0.00075905) and minimal negative correlations with weight (-9.555753e-05) and age (-0.002927279).

F. Comparative Analysis

Ranking the models by MSE from best to worst:

1) Combined Exclamations and Review Length (MSE = 0.49700863)
2) Summary Exclamation Count (MSE = 0.49644174)
3) Full Review Exclamation Count (MSE = 0.49989404)
4) Physical Characteristics (MSE = 0.515731479553)
5) Review Length (MSE = 0.51629136)
6) Fixed Value (MSE = 0.51638951)

This analysis reveals that linguistic features, particularly exclamation marks, are the strongest predictors of user ratings. The combined model's superior performance suggests that considering both the enthusiasm (exclamation marks) and conciseness (review length) of reviews provides the most accurate predictions. Physical characteristics offer some predictive power but are less reliable than review content. These baseline models provide valuable insights for developing more sophisticated predictive algorithms for user satisfaction in clothing rental services.

VI. Developing Precise Models

A. Library Implementation: Latent Factors using SVD

To implement collaborative filtering, we employed Single Value Decomposition (SVD) using the Surprise library. This approach is based on matrix factorization, decomposing the user-item interaction matrix into latent factors that represent users and items.

1) Methodology: We utilized GridSearchCV, a model utility from the Surprise library, for hyperparameter tuning. This method employs grid search combined with cross-validation to find the optimal set of hyperparameters.

a) Cross-Validation Setup:

- Folds: 3-fold cross-validation
- Process: The data was split into 3 subsets, with the model training on 2 subsets and testing on the third, iteratively.

b) Hyperparameters Tuned:

- **n_factors**: Number of latent factors to capture user and item characteristics.
  - Values tested: [1, 3, 5, 10, 15]
  - Significance: Higher values capture more complex interactions but risk overfitting.
- **lr_all**: Learning rate for stochastic gradient descent.
  - Values tested: [0.005, 0.01, 0.02]
  - Significance: Controls the step size during optimization. Lower values may lead to slower convergence but potentially better accuracy.
- **reg_all**: Regularization term to prevent overfitting.
  - Values tested: [0.1, 0.2, 0.3, 0.4]
  - Significance: Higher values increase regularization, potentially reducing overfitting at the cost of model flexibility.
- **n_epochs**: Number of epochs (training iterations).
  - Values tested: [30, 40, 50]
  - Significance: More epochs allow for more training but risk overfitting and increased computational cost.

2) Optimal Hyperparameters: After executing the grid search, the following hyperparameters yielded the best performance:

- n_factors = 1
- reg_all = 0.2
- lr_all = 0.005
- n_epochs = 30

a) Analysis of Optimal Parameters:

- The low value of n_factors (1) suggests that a single dimension of preferences, such as a general tendency to like or dislike items, is sufficient for this dataset. This simplicity may help prevent overfitting.
- The moderate regularization (reg_all = 0.2) balances between model flexibility and overfitting prevention.
- A low learning rate (lr_all = 0.005) indicates that small, cautious steps in optimization were most effective, potentially due to noise or complexity in the data.
- The relatively low number of epochs (30) suggests that the model converges quickly, possibly due to the simplicity of the latent space (n_factors = 1).

3) Model Characteristics:

- **Dimensionality**: The single latent factor (n_factors = 1) implies that the model captures only one dimension of user-item interactions. This could represent an overall preference score or general satisfaction level.
- **Interpretability**: With only one factor, the model is highly interpretable. Each user and item is essentially represented by a single number, which could be seen as their position on a single spectrum of preference.
- **Scalability**: This model is well-suited for large, sparse datasets typical in recommendation systems. The low dimensionality (1 factor) ensures computational efficiency.
- **Generalization**: The combination of low complexity (1 factor), moderate regularization, and a conservative learning rate suggests a model that generalizes well, avoiding overfitting to noise in the training data.

4) Model Performance: After training the SVD model with the optimal hyperparameters, we evaluated its performance on the test set:

- Mean Squared Error (MSE): 0.3687

a) Analysis of MSE:

- The MSE of 0.3687 indicates a strong predictive performance, considering the rating scale of 0-5.
- This error translates to an average deviation of approximately 0.607 (square root of MSE) rating points from the true values.
- Given the simplicity of the model (single latent factor), this low MSE suggests that the model captures a significant portion of the variance in user preferences.
- The performance is comparable to, and slightly better than, the TF-IDF based models (MSE 0.379), indicating that the latent factor approach is effective in this domain.

5) Implications and Future Directions:

- The effectiveness of a single-factor model, achieving an MSE of 0.3687, suggests that user preferences in this clothing rental system might be relatively straightforward, possibly dominated by overall satisfaction rather than complex, multidimensional preferences.
- The strong performance with minimal complexity reinforces the idea that the model generalizes well to unseen data.
- Future work could explore:
  - Incorporating additional features (e.g., item categories, user demographics) into a hybrid model to potentially further reduce the MSE.
  - Experimenting with non-linear models or deep learning approaches to capture more complex patterns and potentially improve upon the current MSE.
  - Investigating time-based effects, such as evolving user preferences or seasonal trends, to enhance the model's predictive power.
  - Comparing this SVD implementation with other collaborative filtering techniques to benchmark its performance in the context of clothing rental recommendations.

This SVD implementation, achieving an MSE of 0.3687, provides a robust baseline for collaborative filtering in the clothing rental recommendation system, balancing simplicity, interpretability, and predictive power. The model's performance suggests that even with a single latent factor, it captures significant patterns in user-item interactions, offering a strong foundation for further refinements and comparisons with other approaches.

B. Term Frequency-Inverse Document Frequency

This section details the implementation and evaluation of a Term Frequency-Inverse Document Frequency (TF-IDF) model for predicting user ratings in a clothing rental recommendation system. The model utilizes a dataset of 192,544 user reviews to extract meaningful features from text and predict user satisfaction.

1) Data Preprocessing and Feature Extraction: The preprocessing pipeline consists of several key steps:

- Loading 192,544 reviews from a gzipped JSON file.
- Splitting the data into training (90%, 173,289 reviews) and test (10%, 19,255 reviews) sets.
- Removing punctuation and converting all text to lowercase to standardize the input.
- Creating a vocabulary of the 1000 most frequent words from the training set, balancing feature richness and computational efficiency.
- Generating a word-to-index mapping for efficient feature vector creation.

2) TF-IDF Matrix Construction: A sparse matrix is constructed to represent the TF-IDF features:

- Each row corresponds to a review (192,544 rows).
- Each column represents a word in the vocabulary (1000 columns) plus an additional column for the bias term (1001 total features).
- The matrix is implemented using SciPy's `lil_matrix` for memory-efficient sparse representation.
- TF-IDF values are implicitly calculated: term frequency is represented by word counts, while IDF is captured by the selection of top 1000 words.

3) Rating Normalization: User ratings undergo normalization:

- Original ratings (0-10 scale) are divided by 2 to transform to a 0-5 scale.
- Missing ratings are imputed with the average rating from the training set.

4) Model Architecture: The prediction system utilizes a linear regression model:

- Implemented using scikit-learn's `LinearRegression` class.
- Trained on the TF-IDF features (x_{train}) and corresponding normalized ratings (y_{train}).
- The model learns coefficients for each word in the vocabulary, capturing the importance of terms in predicting ratings.
- A bias term allows the model to learn a global offset.

5) Training and Evaluation: Two models were trained and evaluated:

- Bag of Words Model (MSE): 0.3793067
- TF-IDF Model (MSE): 0.37930665

The marginal improvement in the second model (difference of 8.5682e-8 in MSE) suggests a slight refinement, possibly due to hyperparameter tuning or minor adjustments in feature engineering.

6) Analysis of Results: The low MSE values (approximately 0.379) for both models indicate strong predictive performance:

- On the normalized 0-5 scale, an MSE of 0.379 translates to an average prediction error of about 0.615 (square root of MSE).
- This suggests that, on average, the model's predictions deviate by about 0.615 points from the actual ratings, demonstrating good accuracy in capturing user satisfaction.
- The similarity between the two models' performance indicates consistency in the approach and suggests that the feature extraction and modeling process is stable.

7) Implications and Future Directions: The TF-IDF based approach proves effective in capturing the relationship between review text and user ratings:

- The model's performance suggests that review text contains valuable information for predicting user satisfaction in clothing rental services.
- Future work could explore:
  - Expanding the vocabulary size to capture more nuanced language patterns.
  - Incorporating advanced NLP techniques like word embeddings or sentiment analysis.
  - Exploring non-linear models to capture complex relationships between text features and ratings.
  - Investigating the impact of additional metadata (e.g., user demographics, item characteristics) on prediction accuracy.

In conclusion, this TF-IDF based approach provides a solid foundation for predicting user satisfaction in clothing rental systems, demonstrating the power of text analysis in understanding and forecasting customer experiences.

C. Custom Implementation: Latent Factor Model

We developed a custom latent factor model to address the specific characteristics of our clothing rental dataset. This approach leverages the strengths of matrix factorization while incorporating strategies to mitigate common challenges.

1) Model Overview: Latent factor models are widely used in recommendation systems due to their ability to model user preferences by learning from sparse user-item interaction matrices. These models learn latent factors for users and items by minimizing an objective function that captures the error between predicted and observed ratings.

2) Advantages of Latent Factor Models:

a) User and Item Interaction Modeling: Our dataset contains 192,544 user-item interactions, representing a clear structure of ratings given by users for rented clothing items. Latent factor models excel at uncovering underlying patterns in such data, potentially capturing preferences for specific styles or fits.

b) Handling Sparsity: The dataset exhibits extreme sparsity:

- Total possible interactions: $105,508 \times 5,850 \approx 617 million$
- Observed interactions: 192,544
- Sparsity ratio: $\frac{192,544}{617,000,000} \approx 0.0312\%$

Latent factor models, particularly matrix factorization techniques, are designed to handle such high levels of sparsity effectively.

c) Scalability: Our implementation uses closed-form solutions to calculate bias and gamma parameters, eliminating the need for gradient-based updates. This approach enhances computational efficiency and convergence speed, crucial for our dataset size.

d) Feature-Agnostic Predictions: The model focuses on learning from user-item rating interactions without requiring additional metadata, making it versatile and potentially more generalizable.

3) Challenges and Mitigations:

a) Highly Skewed Ratings:

- Issue: Ratings are concentrated in the upper range (8-10 on a 0-10 scale).
- Mitigation: We normalized ratings to a 0-5 scale, improving the spread of ratings without loss of information. This resulted in a lower Mean Squared Error (MSE) during training.

b) Loss of Contextual Information:

- Issue: Rich metadata (user demographics, textual reviews) is not directly utilized in traditional latent factor models.
- Future Work: Explore hybrid models that incorporate metadata as additional features or constraints in the latent factor learning process.

c) Cold Start Problem:

- Issue: New users or items lack sufficient interactions for accurate predictions.
- Mitigation: Our approach of considering all sizes of an item as a single product helps build a denser interaction matrix, potentially alleviating some cold start issues.

4) Novel Approaches:

a) Product Definition: Unlike previous work [1] that treated each size of an item as a separate product, we consider all sizes of an item as a single product. This approach offers several advantages:

- Increased density: Reduces the sparsity of the user-item interaction matrix.
- Assumption: Users rate an item once they find a size that fits, allowing the model to focus on overall item satisfaction rather than size-specific ratings.
- Potential improvement: May capture general item preferences more effectively, especially for users who have tried multiple sizes of the same item.

b) Rating Scale Normalization:

- Original scale: 0-10
- Normalized scale: 0-5
- Benefits:
  - Improved rating distribution
  - Lower MSE during training
  - Potentially better differentiation between "good" and "excellent" ratings

5) Model Performance and Implications: Our custom implementation of the latent factor model achieved a Mean Squared Error (MSE) of 0.317 on the test set. This performance metric provides valuable insights:

- Strong predictive power: The MSE of 0.317 indicates high accuracy in predicting user ratings, outperforming both the TF-IDF models ($MSE \approx 0.379$) and the SVD implementation ($MSE \approx 0.3687$).
- Effective sparsity handling: The lower MSE suggests that treating items across sizes as single entities successfully reduces the effective sparsity of the dataset.
- Improved rating differentiation: The normalized rating scale (0-5) appears to enhance the model's ability to distinguish between different levels of user satisfaction.
- Computational efficiency: The use of closed-form solutions for certain parameters likely contributed to both the model's performance and its scalability.

The superior performance of this custom latent factor model, achieving an MSE of 0.317, validates our approach to addressing the dataset's specific challenges. It demonstrates that domain-specific optimizations, such as our novel product definition and rating scale normalization, can significantly enhance the effectiveness of traditional matrix factorization techniques in the context of clothing rental recommendations.

6) Future Directions:

- Hybrid models: Incorporate user and item metadata to enhance prediction accuracy.
- Temporal dynamics: Explore time-based effects on user preferences and item popularity.
- Advanced regularization: Implement techniques like adaptive regularization to further mitigate overfitting and improve generalization.
- Ensemble methods: Combine latent factor models with other approaches (e.g., content-based filtering) for more robust recommendations.

This custom implementation of a latent factor model, with its impressive MSE of 0.317, provides a highly effective and tailored approach to the clothing rental recommendation system. It successfully balances the strengths of traditional matrix factorization with domain-specific optimizations, setting a strong benchmark for future improvements and comparisons with other recommendation techniques in this domain.

VII. Results

This section presents a comprehensive analysis of the various models developed for predicting user ratings in our clothing rental recommendation system. We compare the performance of baseline models, linear regression models, and more advanced techniques such as TF-IDF, SVD, and custom latent factor models.

A. Model Performance Comparison

Table V summarizes the Mean Squared Error (MSE) for each model:

Model	MSE
Fixed Value (Median)	0.51639
Length of Review	0.51629
Count of Exclamation Marks (Full Review)	0.49989
Count of Exclamation Marks (Summary)	0.49644
Combined Exclamations and Review Length	0.49701
Physical Characteristics	0.51573
Term Frequency-Inverse Document Frequency	0.37930
SVD (Surprise Library Implementation)	0.36870
Custom Latent Factor Model	0.31700

TABLE V: Performance Comparison of Different Models

B. Analysis of Results

1) Baseline and Simple Linear Models: The Fixed Value (Median) model, with an MSE of 0.51639, serves as our primary baseline. This model simply predicts the median rating for all instances, providing a benchmark for the minimum acceptable performance. Surprisingly, the Length of Review model (MSE = 0.51629) shows negligible improvement over this baseline, suggesting that review length alone is not a strong predictor of user satisfaction.

2) Linguistic Feature Models: Models based on exclamation marks show a marked improvement over the baseline:

- Full Review Exclamation Count: MSE of 0.49989
- Summary Exclamation Count: MSE of 0.49644
- Combined Exclamations and Review Length: MSE of 0.49701

These results indicate that the presence of exclamation marks, particularly in review summaries, is a stronger predictor of user satisfaction than review length. The combined model's performance suggests that while review length adds some predictive power, it's less influential than the enthusiasm expressed through exclamation marks.

3) Physical Characteristics Model: With an MSE of 0.51573, this model slightly outperforms the baseline but falls short of the linguistic feature models. This indicates that while physical attributes like height, weight, and age have some predictive power, they are less reliable indicators of user satisfaction compared to review content.

4) Advanced Models: The advanced models show significant improvements in predictive power:

a) Term Frequency-Inverse Document Frequency (TF-IDF): With an MSE of 0.37930, the TF-IDF model demonstrates a substantial improvement over simpler approaches. This indicates that the content of reviews, when properly weighted for term importance, provides valuable insights into user satisfaction.

b) SVD (Surprise Library Implementation): The SVD model achieves an MSE of 0.36870, slightly outperforming the TF-IDF approach. This suggests that latent factor models can effectively capture underlying patterns in user-item interactions, even with a single latent factor.

c) Custom Latent Factor Model: Our custom latent factor model achieves the best performance with an MSE of 0.31700. This significant improvement over other approaches validates our strategies for handling data sparsity, rating normalization, and computational efficiency.

C. Comparative Analysis

1) Simple linguistic features (exclamation marks) outperform baseline and physical characteristic models, highlighting the importance of sentiment and enthusiasm in predicting user satisfaction.

2) The TF-IDF model's strong performance (MSE reduction of 26.5% compared to the best simple model) underscores the value of considering the full content of reviews, not just simple features.
3) The SVD model's performance, with a 4.5% improvement over TF-IDF, demonstrates the power of latent factor approaches in capturing complex user-item interactions.
4) Our custom latent factor model shows the most substantial improvement, with a 38.5% reduction in MSE compared to the best simple model and a 14% improvement over the SVD approach. This highlights the benefits of tailoring the model to the specific characteristics of the dataset.
5) The progression from simple models ($MSE \approx 0.51$) to advanced techniques (MSE = 0.317) represents a 38% overall improvement in predictive accuracy, showcasing the potential of sophisticated modeling approaches in this domain.

D. Implications and Future Directions

- The strong performance of linguistic feature models suggests that incorporating more sophisticated natural language processing techniques could yield further improvements.
- The superior performance of latent factor models indicates that capturing underlying patterns in user-item interactions is crucial for accurate predictions in this domain.
- The relative underperformance of physical characteristic models suggests that user satisfaction might be more dependent on subjective experience than objective fit metrics.
- Future work should focus on:
  - Developing hybrid models that combine the strengths of content-based approaches (like TF-IDF) with collaborative filtering techniques (like SVD).
  - Exploring more complex latent factor models, potentially incorporating temporal dynamics to capture evolving user preferences.
  - Investigating the integration of additional metadata (e.g., item categories, user demographics) into the custom latent factor model to potentially enhance its predictive power further.
  - Developing ensemble methods that leverage the strengths of multiple model types, potentially combining the insights from linguistic, latent factor, and metadata-based approaches.

VIII. Conclusion

This study has explored various approaches to predicting user ratings in a clothing rental recommendation system, progressing from simple baseline models to sophisticated machine learning techniques. Our analysis reveals a clear improvement in predictive accuracy as we move from basic models to more advanced approaches. The custom latent factor model emerged as the most effective, achieving an MSE of 0.317, significantly outperforming other methods including TF-IDF ($MSE \approx 0.37930$) and SVD ($MSE \approx 0.36870$). This model's success can be attributed to its ability to capture complex user-item interactions while addressing domain-specific challenges such as data sparsity and rating normalization.

A key outcome of this research is the development of a practical recommender system capable of generating personalized item recommendations for individual users. This system can predict the top 5 items a user is likely to rent, along with estimated ratings. For example:

```
Top 5 recommendations for user 420272:
Item: 1859039, Estimated Rating: 10.00
Item: 182915, Estimated Rating: 10.00
Item: 123793, Estimated Rating: 10.00
Item: 2057975, Estimated Rating: 10.00
Item: 134393, Estimated Rating: 10.00
```

This functionality demonstrates the model's potential for real-world application in e-commerce and rental platforms. By leveraging users' previous interactions, the system can predict future preferences with high accuracy, potentially enhancing user experience and driving business growth. The success of our custom latent factor model in generating these precise recommendations underscores the value of tailored approaches in recommendation systems. It shows that by carefully addressing domain-specific challenges and leveraging the strengths of matrix factorization techniques, we can create highly effective predictive models.

Future work could focus on incorporating more diverse data sources, exploring temporal dynamics in user preferences, and investigating the integration of this recommender system into a full-fledged e-commerce platform. Additionally, real-world testing and user feedback would be crucial to further validate and refine the model's performance. In conclusion, this study not only advances our understanding of recommendation systems in the clothing rental domain but also provides a practical tool with immediate applications in personalized marketing and user experience enhancement.

REFERENCES

[1] Rishabh Misra, Mengting Wan, and Julian McAuley. "Decomposing fit semantics for product size recommendation in metric spaces". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys '18. Vancouver, British Columbia, Canada: Association for Computing Machinery, 2018, 422–426.

[2] Yifan Hu, Yehuda Koren, and Chris Volinsky. "Collaborative filtering for implicit feedback datasets". In: *IEEE Transactions on Knowledge and Data Engineering* 22.8 (2008), pp. 1263–1284.

[3] Wen Chen et al. "POG: Personalized outfit generation for fashion recommendation at Alibaba iFashion". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2662–2670.

[4] Anonymous. *Enhanced Content-Based Fashion Recommendation System Using Deep Ensemble Classifiers*. Unpublished manuscript. 2024.

[5] Xiaoling Gu et al. "Understanding fashion trends from street photos via neighbor-constrained embedding learning". In: *Proceedings of the 25th ACM international conference on Multimedia*. 2017, pp. 190–198.

[6] Sebastian Heinz, Christian Bracher, and Roland Vollgraf. "LSTM-based dynamic customer model for fashion recommendation". In: *arXiv preprint arXiv:1708.07347* (2017).

[7] Anonymous. *A Comprehensive Framework for Evaluating Recommendation Systems*. Unpublished manuscript. 2024.

[8] Anonymous. *Addressing Challenges in Fashion Rental Recommendation Systems*. Unpublished manuscript. 2024.