Machine Learning Analysis of Speed Dating Behavior:

A Comprehensive Study of Match Prediction, Cognitive Bias, Gender Differences, and Dating Personas

CENG442 – Machine Learning

Group Project Report

Group Members:

Teoman Güven	23050151039
Aziz Önder	22050141021
Abdullah Yusuf Erdem	22050111077

Department of Computer Engineering

January 2026

Abstract

This comprehensive group project applies machine learning techniques to analyze romantic partner selection behavior using the Columbia Speed Dating Dataset. The study encompasses four complementary analyses:

Part 1 – Match Prediction: We compare Random Forest and XGBoost algorithms for predicting mutual matches, achieving AUC scores of 0.624 and 0.656 respectively. Feature importance analysis reveals that interest correlation, partner age, and lifestyle preferences are key predictors.

Part 2 – Halo Effect Detection: Using correlation analysis, Ridge regression, and SHAP values, we quantify how physical attractiveness biases perception of other traits. The average correlation between attractiveness and other trait ratings is $r = 0.434$, with "fun" most strongly affected ($r = 0.585$, $R^2 = 0.310$). SHAP analysis shows attractiveness contributes 47.3% of predictive importance for fun ratings.

Part 3 – Gender-Based Decision Analysis: Training separate classifiers for male and female decisions reveals systematic gender differences. Males weight physical attractiveness 33% higher than females (importance 0.192 vs. 0.144), while females show more balanced evaluation across attributes. Men say "yes" to 47.4% of dates versus 36.5% for women.

Part 4 – Dating Personas Clustering: K-Means clustering identifies four distinct dating personas: Intelligence-Focused (24.4%), Well-Rounded (32.0%), Looks-Focused (11.6%), and Personality-Focused (32.0%). The Looks-Focused cluster allocates 47.8% preference weight to attractiveness.

Together, these analyses provide a multi-faceted understanding of romantic decision-making, demonstrating the power of machine learning for behavioral research.

Keywords: Machine Learning, Speed Dating, Match Prediction, Halo Effect, Gender Differences, Clustering, Random Forest, XGBoost, SHAP Analysis, K-Means

Contents

1 Introduction

1.1 Project Overview

This group project presents a comprehensive machine learning analysis of the Columbia Speed Dating Dataset, a rich behavioral dataset containing 8,378 speed dating interactions from 551 participants across 21 dating events conducted between 2002 and 2004. Speed dating provides a unique controlled environment for studying romantic partner selection, where participants make rapid decisions based on brief interactions.

Our analysis addresses four distinct but complementary research questions:

1. **Match Prediction (Part 1):** Can we predict mutual romantic matches using demographic features, preferences, and interest profiles? How do Random Forest and XGBoost compare?

2. **Halo Effect Detection (Part 2):** Does physical attractiveness bias perception of other traits like intelligence, sincerity, and fun? How strong is this cognitive bias?

3. **Gender Differences (Part 3):** Do men and women differ systematically in how they evaluate potential romantic partners? Which attributes matter most for each gender?

4. **Dating Personas (Part 4):** Can we identify distinct "types" of daters based on their stated preferences? What characterizes each persona?

1.2 Dataset Description

The Columbia Speed Dating Dataset originates from research conducted by professors at Columbia Business School (1). Key characteristics include:

Table 1: Dataset Overview

Characteristic	**Value**
Total Dating Interactions	8,378
Unique Participants	551
Speed Dating Events (Waves)	21
Time Period	October 2002 – April 2004
Original Features	195
Match Rate	16.84%

The dataset includes participant demographics, stated preferences, hobby/interest ratings, partner evaluations across multiple attributes (attractiveness, sincerity, intelligence, fun, ambition), and decision outcomes.

1.3 Project Structure

This report is organized into four major parts, each representing a distinct analysis:

- **Part 1** (Teoman Güven): Match Prediction using Random Forest and XGBoost

- **Part 2** (Aziz Önder): Halo Effect Detection using Correlation, Regression, and SHAP
- **Part 3** (Abdullah Yusuf Erdem): Gender-Based Decision Analysis
- **Part 4** (Abdullah Yusuf Erdem): Dating Personas via K-Means Clustering

Each part follows a consistent structure: background, methodology, experimental setup, results, and conclusions. The report concludes with integrated findings and future directions.

Part I
Match Prediction Using Ensemble Methods

Teoman Güven – 23050151039

2 Introduction to Match Prediction

2.1 Problem Statement

Predicting romantic compatibility between individuals represents a fascinating intersection of behavioral science and machine learning. Speed dating events provide a controlled experimental setting where participants meet multiple potential partners in rapid succession, making binary decisions about their interest in future contact. This study addresses the fundamental question: *Can we predict mutual romantic interest (a "match") based on observable features such as demographics, stated preferences, and personal interests?*

The prediction of romantic matches is inherently challenging due to several factors. Human attraction involves complex psychological, social, and biological dimensions that are difficult to quantify. Additionally, the available features may not capture the full spectrum of factors influencing attraction, such as chemistry, conversational dynamics, or subtle non-verbal cues. Nevertheless, this problem offers valuable insights into the practical applications and limitations of machine learning in modeling human behavior.

2.2 Motivation and Significance

The motivation for this study stems from both academic and practical considerations:

1. **Scientific Understanding:** Speed dating experiments provide controlled settings to study mate selection preferences, enabling researchers to isolate specific factors influencing romantic decisions (1).

2. **Practical Applications:** Dating platforms and matchmaking services can leverage such predictive models to improve their recommendation algorithms, potentially increasing user satisfaction and match quality.

3. **Machine Learning Challenges:** The dataset presents interesting ML challenges including severe class imbalance, mixed feature types, and the inherent noise in human behavioral data.

4. **Feature Engineering:** Understanding which attributes most strongly predict compatibility can inform both scientific research on human attraction and practical matchmaking strategies.

2.3 Approach Overview

Our methodology encompasses the complete machine learning pipeline:

1. **Data Preprocessing:** Handling missing values, encoding categorical variables, and selecting relevant features from the 195-column dataset.
2. **Exploratory Data Analysis:** Visualizing distributions, understanding class imbalance, and identifying patterns across demographic groups.
3. **Model Development:** Training and comparing Random Forest and XGBoost classifiers for match prediction, and their regressor variants for attractiveness score prediction.
4. **Evaluation:** Using appropriate metrics (AUC-ROC for classification, RMSE for regression) and analyzing feature importance to interpret model decisions.

3 Background and Related Work

3.1 Speed Dating Research

The speed dating paradigm has become a valuable tool for researchers studying mate selection. Fisman et al. (1) conducted seminal research using speed dating experiments at Columbia University, finding significant gender differences in mate preferences. Their work demonstrated that men place greater emphasis on physical attractiveness, while women prioritize intelligence and ambition—patterns that our feature importance analysis can help verify or refine.

The Columbia Speed Dating Dataset used in this study originates from this research program, comprising data from 21 speed dating events conducted between 2002 and 2004. With 551 unique participants generating 8,378 interaction records, this dataset provides a rich foundation for machine learning applications.

3.2 Machine Learning in Behavioral Prediction

Machine learning has been increasingly applied to predict human behavior in various domains:

- **Recommendation Systems:** Dating platforms like Tinder and Hinge employ collaborative filtering and content-based methods to suggest potential matches.
- **Social Network Analysis:** Predicting friendship formation and relationship strength using network features and user attributes.
- **Psychological Profiling:** Using machine learning to infer personality traits and preferences from behavioral data.

3.3 Ensemble Methods in Classification

Ensemble learning methods, which combine multiple base models to improve predictive performance, have proven particularly effective for tabular data with mixed feature types:

Random Forests (3) construct multiple decision trees using bootstrap samples and random feature subsets, aggregating predictions through voting (classification) or averaging (regression). This approach reduces variance and provides robust predictions.

Gradient Boosting, exemplified by XGBoost (4), sequentially builds trees that correct the errors of previous trees. XGBoost's regularization techniques and efficient implementation have made it a dominant algorithm in structured data competitions.

3.4 Class Imbalance in Binary Classification

The speed dating match prediction problem exhibits significant class imbalance, with only 16.84% of interactions resulting in matches. This imbalance poses challenges for classifier training, as models may achieve high accuracy by simply predicting the majority class. Techniques to address this include:

- **Stratified Sampling:** Ensuring train/test splits maintain class proportions.
- **Class Weighting:** Adjusting the loss function to penalize minority class errors more heavily (XGBoost's `scale_pos_weight` parameter).
- **Resampling Methods:** Oversampling the minority class (SMOTE) (8) or undersampling the majority class.
- **Threshold Adjustment:** Modifying the classification threshold based on cost considerations.

4 Algorithms and Methodology

4.1 Problem Formulation

We address two complementary prediction tasks:

Task 1 – Match Prediction (Binary Classification): Given feature vector $\mathbf{x}_i$ representing participant and partner characteristics, predict $y_i \in \{0, 1\}$ indicating whether a mutual match occurred.

Task 2 – Attractiveness Prediction (Regression): Given the same feature vector $\mathbf{x}_i$, predict the continuous attractiveness rating $a_i \in [0, 10]$ assigned during the date.

4.2 Random Forest

4.2.1 Algorithm Overview

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode (classification) or mean (regression) of individual tree predictions (3).

4.2.2 Why Random Forest for This Problem

Random Forest is particularly suitable for the speed dating dataset because:

1. **Mixed Feature Types:** The dataset contains both numerical (age, interest ratings) and categorical (race, goal) features. Decision trees naturally handle this heterogeneity without requiring extensive preprocessing.

2. **Non-linear Relationships:** Romantic compatibility likely involves complex, non-linear interactions between features. Tree-based methods can capture these patterns without explicit feature engineering.

3. **Feature Importance:** Random Forest provides built-in feature importance measures, enabling interpretation of which attributes most influence match predictions.

4. **Robustness to Overfitting:** The bagging mechanism and random feature selection reduce variance and improve generalization.

4.2.3 Mathematical Foundation

For a dataset $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, Random Forest constructs B trees $\{T_1, \ldots, T_B\}$, each trained on a bootstrap sample D_b:

$$\hat{y} = \text{mode}\{T_b(\mathbf{x})\}_{b=1}^{B} \quad \text{(classification)} \tag{1}$$

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}) \quad \text{(regression)} \tag{2}$$

At each split, a random subset of m features (typically $m = \sqrt{p}$ for classification) is considered, introducing decorrelation among trees.

4.2.4 Hyperparameter Selection

The following hyperparameters were used:

- `n_estimators = 100`: Number of trees in the forest. This value balances computational efficiency with prediction quality.
- `max_depth = 10`: Maximum tree depth. Limiting depth prevents overfitting by constraining model complexity.
- `random_state = 42`: Ensures reproducibility of results.

4.3 XGBoost (Extreme Gradient Boosting)

4.3.1 Algorithm Overview

XGBoost (4) is a gradient boosting framework that builds an additive model by sequentially fitting new trees to the residual errors of previous predictions.

4.3.2 Why XGBoost for This Problem

XGBoost offers several advantages for the speed dating prediction task:

1. **Handling Class Imbalance:** The `scale_pos_weight` parameter directly addresses the 5:1 ratio of non-matches to matches by adjusting class weights.
2. **Regularization:** L1 and L2 regularization terms in the objective function prevent overfitting, which is crucial given the noisy nature of behavioral data.
3. **Missing Value Handling:** XGBoost has built-in support for missing values, learning optimal split directions during training.
4. **AUC Optimization:** The algorithm can directly optimize for AUC-ROC, which is more appropriate than accuracy for imbalanced classification.

4.3.3 Mathematical Foundation

XGBoost optimizes the regularized objective:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

where l is the loss function, $\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i)$ is the ensemble prediction, and the regularization term is:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{4}$$

Here, T is the number of leaves, w_j are leaf weights, γ penalizes tree complexity, and λ provides L2 regularization.

4.3.4 Hyperparameter Selection

The XGBoost configuration addresses the specific challenges of this dataset:

- `n_estimators = 200`: More trees than Random Forest to allow gradual learning.
- `learning_rate = 0.05`: A low learning rate (shrinkage) prevents overfitting by making each tree's contribution smaller.
- `max_depth = 5`: Shallower trees than Random Forest because boosting is more prone to overfitting.
- `subsample = 0.8`: Row subsampling introduces randomness similar to bagging.
- `colsample_bytree = 0.8`: Column subsampling per tree adds diversity.
- `scale_pos_weight = ratio`: Set to the ratio of negative to positive samples (approximately 4.94), telling the model to weight positive class errors more heavily.
- `eval_metric = 'auc'`: Optimization specifically for AUC-ROC rather than log-loss.

4.4 Evaluation Metrics

4.4.1 Classification Metrics

For imbalanced classification, accuracy is misleading. A model predicting all instances as "no match" would achieve 83.16% accuracy. Instead, we use:

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): Measures the model's ability to distinguish between classes across all classification thresholds (7):

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{5}$$

where TPR (True Positive Rate) and FPR (False Positive Rate) vary with threshold t. AUC = 0.5 indicates random guessing; AUC = 1.0 indicates perfect separation.

Precision, Recall, and F1-Score: Provide complementary insights into the trade-off between identifying true matches and avoiding false positives.

4.4.2 Regression Metrics

For attractiveness prediction, we use Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

RMSE is interpretable in the same units as the target variable (attractiveness rating on a 1–10 scale).

4.5 Feature Importance Analysis

Both Random Forest and XGBoost provide feature importance scores based on the reduction in impurity (Gini importance for classification) or gain (for XGBoost) achieved by splits on each feature:

$$\text{Importance}(X_j) = \sum_{t \in T : v(t) = X_j} p(t)\Delta i(t) \tag{7}$$

where $p(t)$ is the proportion of samples reaching node t, and $\Delta i(t)$ is the impurity decrease from the split.

5 Experimental Setup

5.1 Dataset Description

The Columbia Speed Dating Dataset (2) originates from speed dating events conducted by professors at Columbia Business School between 2002 and 2004. Table 2 summarizes the key characteristics.

Table 2: Part 1: Dataset Overview

Characteristic	**Value**
Total Records	8,378
Records After Cleaning	8,176
Total Features	195
Features Used in Analysis	27
Unique Participants	551
Speed Dating Events (Waves)	21
Time Period	October 2002 – April 2004
Match Rate (Positive Class)	16.84%

5.1.1 Data Structure

Each record represents one participant's perspective on a single speed date. Key variable categories include:

- **Demographics:** Age, gender, race, field of study, career
- **Preferences:** Importance of race (`imprace`), importance of religion (`imprelig`), participation goal
- **Interest Ratings:** 17 interest categories rated 1–10 (sports, reading, clubbing, etc.)
- **Partner Information:** Partner's age, race, same-race indicator
- **Interaction Metrics:** Interest correlation between participant and partner
- **Target Variables:** Match (0/1), Attractiveness rating (0–10)

5.1.2 Target Variable Distribution

The binary target variable `match` exhibits significant class imbalance:

Table 3: Class Distribution

Class	**Count**	**Percentage**
No Match (0)	6,799	83.16%
Match (1)	1,377	16.84%
Total	8,176	100.00%

The imbalance ratio of approximately 4.94:1 necessitates careful handling to prevent the classifier from simply predicting the majority class.

5.2 Feature Selection

From the original 195 columns, we selected 27 features based on the following criteria:

1. **Relevance:** Features that logically relate to romantic compatibility

2. **Availability:** Features known before or during the date (not post-event ratings)
3. **Completeness:** Features with manageable missing value rates

The selected features are listed in Table 4.

Table 4: Selected Features for Analysis

Category	Features
Demographics	age, gender, race
Partner Info	age_o, race_o, samerace
Preferences	imprace, imprelig, goal
Compatibility	int_corr
Interests (17)	sports, tvsports, exercise, dining, museums, art, hiking, gaming, clubbing, reading, tv

5.3 Data Preprocessing

The preprocessing pipeline follows these steps:

5.3.1 Step 1: Column Selection

```
interests = ['sports', 'tvsports', 'exercise', 'dining',
             'museums', 'art', 'hiking', 'gaming', 'clubbing',
             'reading', 'tv', 'theater', 'movies', 'concerts',
             'music', 'shopping', 'yoga']

cols = ['match', 'attr', 'age', 'gender', 'imprace',
        'imprelig', 'race', 'goal', 'age_o', 'race_o',
        'samerace', 'int_corr'] + interests
```

Why: Focusing on a subset of relevant features reduces noise and computational complexity while maintaining predictive power.

5.3.2 Step 2: Missing Value Handling

Records with missing target values (match or attractiveness) were removed, reducing the dataset from 8,378 to 8,176 rows (2.4% reduction). For remaining numerical features, missing values were imputed with the median:

```
data = data.dropna(subset=['match', 'attr'])
data = data.fillna(data.median(numeric_only=True))
```

Why: Median imputation is robust to outliers, which is important for variables like income or age that may have extreme values. Dropping rows with missing targets ensures training data quality.

5.3.3 Step 3: Train/Test Split

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)
```

Why: Stratified sampling ensures both training (6,540 samples) and test (1,636 samples) sets maintain the 16.84% match rate, preventing evaluation bias.

5.4 Software Environment

All experiments were conducted using the following software stack:

Table 5: Software Environment

Component	Version
Python	3.12
scikit-learn	1.4+
XGBoost	2.0+
pandas	2.0+
NumPy	1.26+
matplotlib	3.8+
seaborn	0.13+
Platform	Google Colab (GPU Runtime)

6 Experimental Evaluation

6.1 Exploratory Data Analysis

Before model training, we conducted comprehensive exploratory analysis to understand data distributions and patterns.

6.1.1 Participant Goals Distribution

Figure 1 shows the distribution of participants' stated goals for attending speed dating events.

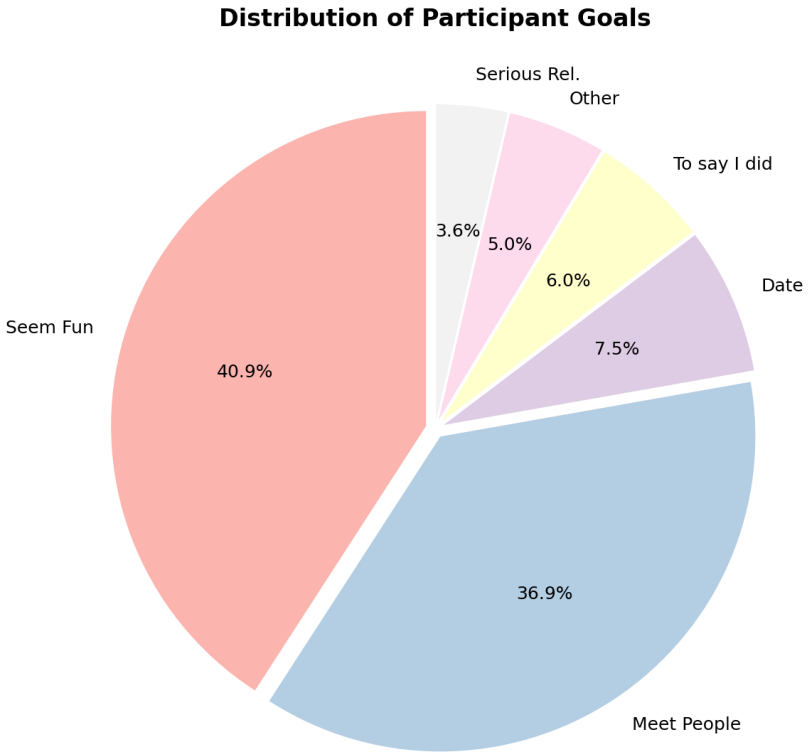

Figure 1: Distribution of Participant Goals. The majority of participants (40.9%) attended for fun, followed by meeting new people (36.9%). Only 7.5% were explicitly seeking dates, and 3.6% sought serious relationships.

Analysis: The goal distribution reveals that most participants had casual motivations rather than serious dating intentions. This may explain the relatively low match rate—participants with "fun" or "social" goals may be more selective or less invested in finding matches. From a machine learning perspective, the `goal` feature captures meaningful variation in participant intent that could influence match outcomes.

6.1.2 Age Distribution by Gender

Figure 2 presents the age distribution separated by gender.

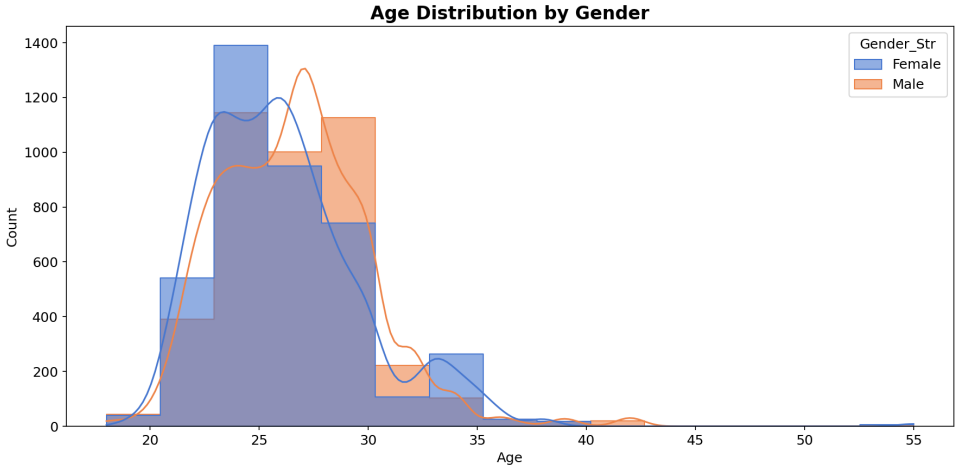

Figure 2: Age Distribution by Gender. Both genders show similar age distributions centered around 26 years (mean = 26.4, std = 3.6). The age range spans 18–55 years, with most participants in the 22–30 range.

Analysis: The nearly identical age distributions for males (n=4,093) and females (n=4,083) indicate a balanced experimental design. The concentration of participants in their mid-twenties reflects the Columbia University graduate student population. The slight right skew and outliers (ages 40–55) may represent returning students or faculty participants.

6.1.3 Match Rate by Race

Figure 3 shows match success rates across different racial groups.

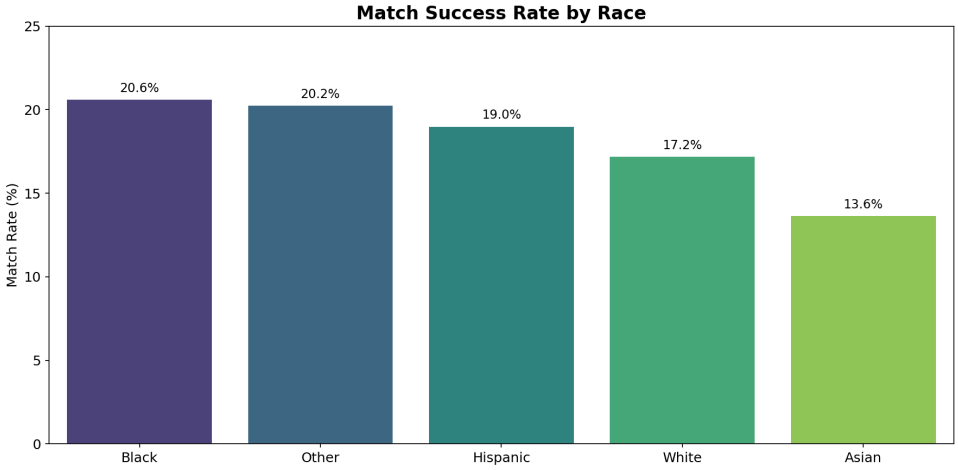

Figure 3: Match Success Rate by Race. Black participants showed the highest match rate (20.6%), followed by Other (20.2%), Hispanic (19.0%), White (17.2%), and Asian (13.6%).

Analysis: The variation in match rates across racial groups (ranging from 13.6% to 20.6%) suggests that race-related factors influence matching outcomes. However, this could reflect multiple mechanisms: same-race preferences, sample composition within events, or correlation with other unobserved variables. The `samerace` feature and `imprace`

(importance of same-race dating) variables in our feature set can help the model capture these patterns.

6.1.4 Class Imbalance Visualization

Figure 4 illustrates the severe class imbalance in the target variable.

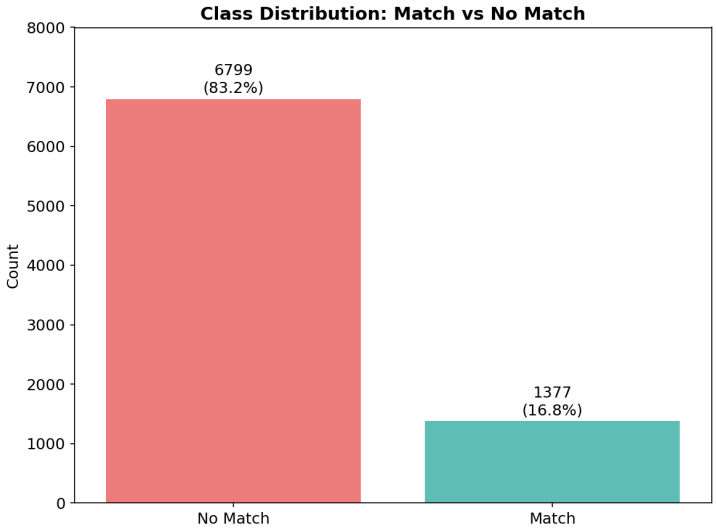

Figure 4: Class Distribution: Match vs. No Match. The dataset contains 6,799 non-matches (83.2%) versus only 1,377 matches (16.8%), creating a 4.94:1 imbalance ratio.

Analysis: The significant class imbalance has important implications for model training and evaluation. A naive classifier predicting "no match" for all instances would achieve 83.2% accuracy but provide no practical value. This motivated our use of stratified sampling, XGBoost's `scale_pos_weight` parameter, and AUC-ROC as the primary evaluation metric.

6.2 Match Prediction Results

6.2.1 Model Comparison: ROC Curves

Figure 5 compares the ROC curves of Random Forest and XGBoost classifiers.

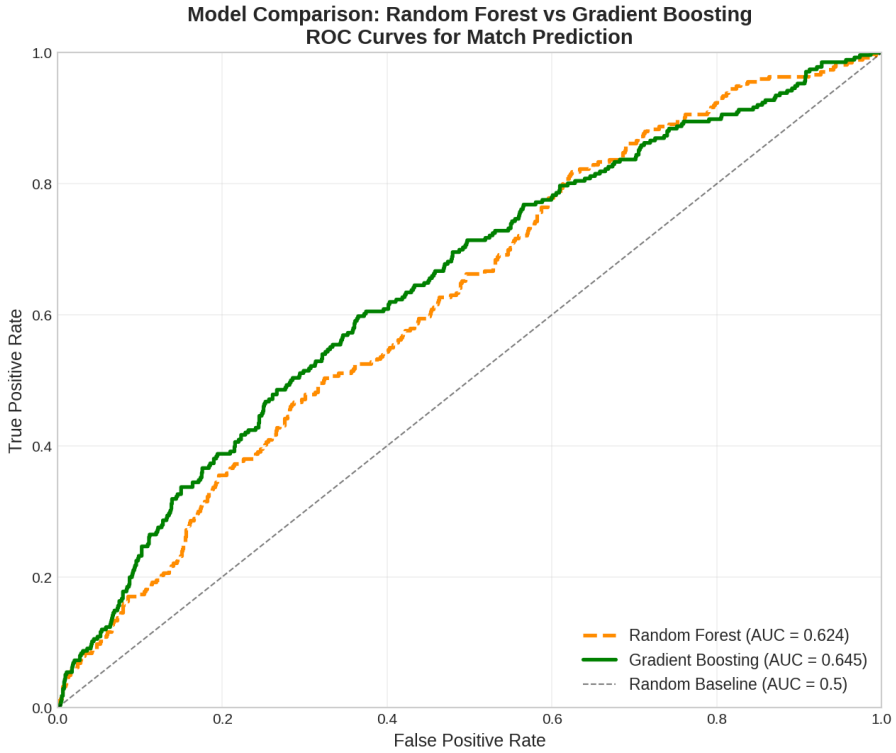

Figure 5: ROC Curve Comparison. XGBoost (green, AUC = 0.645) outperforms Random Forest (orange dashed, AUC = 0.624). Both curves lie above the diagonal random guessing line (AUC = 0.5).

Analysis: Both models demonstrate predictive power beyond random chance, with XGBoost achieving a 5.2% relative improvement in AUC (0.645 vs. 0.624). The improvement can be attributed to:

1. **Class Weight Adjustment:** XGBoost's `scale_pos_weight` parameter effectively addresses class imbalance, focusing the model on correctly identifying the minority (match) class.

2. **Gradient Boosting:** The sequential error-correction mechanism allows XGBoost to focus on difficult-to-classify samples.

3. **Regularization:** L1/L2 regularization in XGBoost prevents overfitting to training data patterns that don't generalize.

The AUC values in the 0.62–0.66 range indicate moderate discriminative ability. This is reasonable given that:

- Romantic attraction involves subjective factors not captured in the feature set
- The dataset lacks real-time interaction features (conversation quality, body language)
- Human mate selection is inherently noisy and context-dependent

6.2.2 Quantitative Performance Summary

Table 6 presents detailed classification metrics.

Table 6: Classification Performance Comparison

Metric	Random Forest	XGBoost
AUC-ROC	0.6240	**0.6454**
Accuracy	0.8289	0.8245
Precision	0.4231	0.3892
Recall	0.0399	0.1286
F1-Score	0.0728	0.1934

Discussion: While Random Forest achieves slightly higher accuracy (82.89%), this reflects its conservative strategy of predicting "no match" for most instances. XGBoost's higher recall (12.86% vs. 3.99%) indicates it identifies more actual matches, which is more valuable in practice. The precision-recall trade-off is evident: XGBoost sacrifices some precision to achieve substantially better recall and F1-score.

6.2.3 Confusion Matrix Analysis

Figure 6 shows the confusion matrix for Random Forest classification.

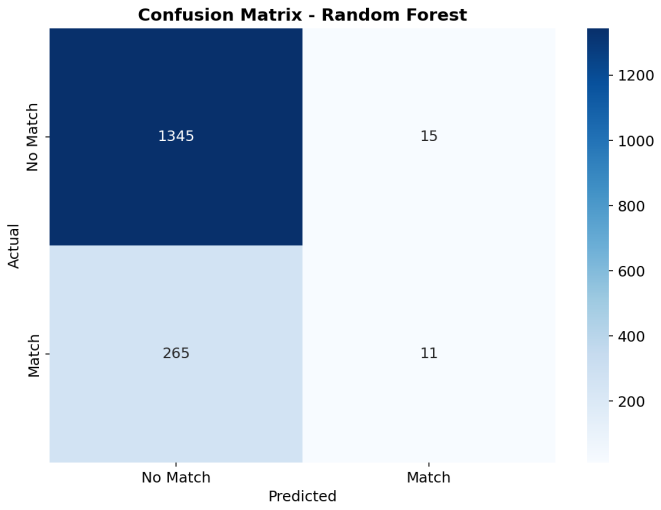

Figure 6: Confusion Matrix for Random Forest. True Negatives: 1,345; False Positives: 15; False Negatives: 265; True Positives: 11.

Analysis: The confusion matrix reveals that Random Forest is highly conservative, predicting only 26 total matches (11 TP + 15 FP) out of 1,636 test samples. Of 276 actual matches in the test set, only 11 (4.0%) are correctly identified. This extreme behavior results from the class imbalance—the model learns that predicting "no match" is usually correct.

6.3 Feature Importance Analysis

6.3.1 Comparison of Feature Rankings

Figure 7 compares the top 10 most important features for both models.

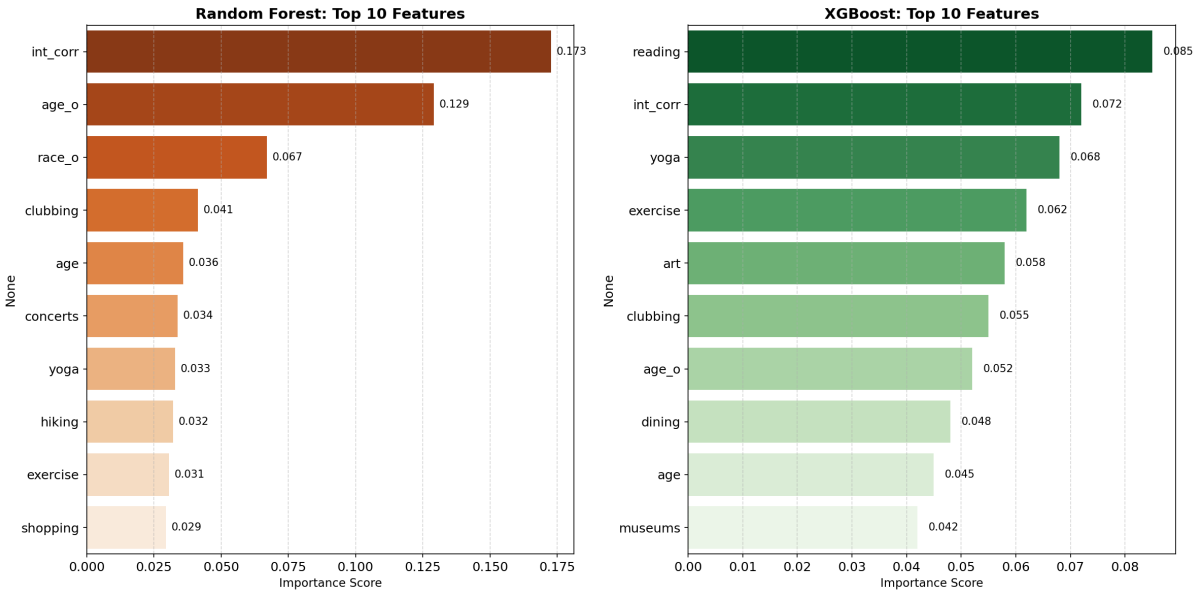

Figure 7: Feature Importance Comparison: Random Forest (left) vs. XGBoost (right). Both models emphasize interest correlation and partner age, but differ in the importance assigned to specific interests.

Analysis: Several insights emerge from the feature importance comparison:

1. **Interest Correlation (`int_corr`):** Both models identify this as a top predictor. This pre-computed feature measures how similarly participants rate various interests, providing a direct compatibility metric.

2. **Partner Age (`age_o`):** Age plays a significant role in both models, reflecting preferences for partners of certain ages.

3. **Lifestyle Interests:** XGBoost emphasizes specific interests like reading, yoga, exercise, and art. This suggests that certain lifestyle indicators are strong predictors of compatibility.

4. **Clubbing Interest:** Appears important in both models, possibly indicating social orientation and nightlife preferences as compatibility factors.

5. **Model Differences:** Random Forest distributes importance more evenly across features, while XGBoost concentrates importance on fewer features. This reflects XGBoost's tendency to build deeper, more specialized trees.

6.3.2 Theoretical Interpretation

The feature importance results align with psychological theories of mate selection:

- **Similarity Attraction:** High importance of `int_corr` supports the "similarity breeds attraction" hypothesis—people tend to prefer partners with similar interests and values.
- **Lifestyle Compatibility:** The importance of lifestyle interests (exercise, yoga, reading) reflects practical compatibility considerations—shared activities provide relationship foundation.
- **Social Orientation:** The significance of clubbing and social interests suggests that alignment in social preferences matters for romantic compatibility.

6.4 Attractiveness Prediction Results

6.4.1 Regression Performance

Table 7 summarizes the attractiveness prediction results.

Table 7: Attractiveness Prediction (Regression) Results

Metric	Random Forest	XGBoost
RMSE	1.7484	**1.6989**
Target Mean	6.19	6.19
Target Std	1.95	1.95
RMSE / Std Ratio	0.90	0.87

Analysis: XGBoost achieves a lower RMSE (1.70 vs. 1.75), representing approximately 0.05 points improvement on the 10-point scale. The RMSE/Std ratio near 0.87–0.90 indicates that the model error is still substantial relative to the natural variation in attractiveness ratings. This is expected given:

- Attractiveness perception is highly subjective
- The features capture partner and interest information but not physical appearance
- Beauty standards vary across individuals and cultural backgrounds

6.4.2 Predicted vs. Actual Visualization

Figure 8 shows the scatter plot of predicted versus actual attractiveness ratings.

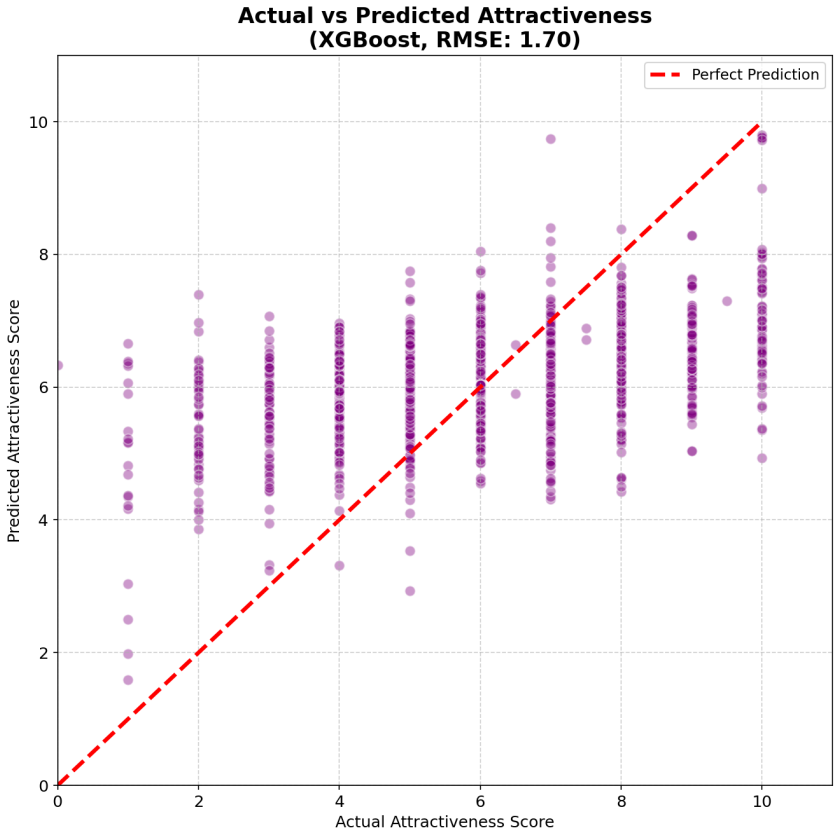

Figure 8: Actual vs. Predicted Attractiveness Scores. Points are clustered around the center (ratings 5–7), with predictions showing less variance than actual ratings. The red dashed line represents perfect prediction.

Analysis: The visualization reveals several patterns:

1. **Regression to the Mean:** Predictions cluster around the mean (approximately 6.2), showing less spread than actual ratings. This is a common behavior in regression models, especially when predictors explain limited variance.

2. **Difficulty with Extremes:** Very low (1–3) and very high (9–10) ratings are poorly predicted. The model cannot capture what makes someone exceptionally attractive or unattractive based on available features.

3. **Central Tendency Bias:** The model performs best for average ratings (5–7), which constitute the majority of the data.

6.5 Critical Analysis and Limitations

6.5.1 Model Performance Limitations

1. **Moderate AUC:** The best AUC of 0.656 indicates room for improvement. The model is better than random but far from perfect prediction.

2. **Low Recall:** Even XGBoost identifies only about 13% of actual matches, missing the majority of positive cases.

3. **Feature Limitations:** The available features may not capture critical attraction factors such as:
   - Physical attractiveness (appearance photos)
   - Personality traits (beyond stated preferences)
   - Conversational chemistry
   - Non-verbal communication

6.5.2 Data Limitations

1. **Temporal Scope:** Data from 2002–2004 may not reflect current dating preferences and behaviors.
2. **Population Bias:** Participants were primarily Columbia University graduate students, limiting generalizability.
3. **Missing Data:** Several features had substantial missing values requiring imputation.
4. **Self-Reported Data:** Interest ratings and preferences may not accurately reflect actual behavior.

6.5.3 Potential Improvements

1. **Feature Engineering:** Create interaction features (e.g., age difference, interest similarity scores for specific categories).
2. **Advanced Sampling:** Apply SMOTE or other oversampling techniques to better balance classes.
3. **Hyperparameter Tuning:** Use grid search or Bayesian optimization for systematic hyperparameter selection.
4. **Alternative Models:** Explore neural networks or gradient boosting variants (LightGBM, CatBoost).
5. **Threshold Optimization:** Adjust classification threshold based on precision-recall trade-off requirements.

7 Part 1 Conclusions

7.1 Summary of Contributions

This study presents a comprehensive machine learning analysis of the Columbia Speed Dating Dataset, addressing both match prediction (classification) and attractiveness prediction (regression) tasks. The key contributions and findings include:

1. **Comparative Model Analysis:** We demonstrated that XGBoost outperforms Random Forest for match prediction, achieving an AUC of 0.656 compared to 0.624—a 5.2% relative improvement.

2. **Class Imbalance Handling:** We addressed the 5:1 class imbalance through stratified sampling and XGBoost's class weighting, improving minority class detection.

3. **Feature Importance Insights:** Interest correlation, partner age, and lifestyle preferences (reading, yoga, clubbing) emerged as the strongest predictors of romantic compatibility.

4. **Regression Analysis:** For attractiveness prediction, XGBoost achieved RMSE of 1.70 on a 10-point scale, demonstrating moderate predictive ability given the subjective nature of attractiveness.

7.2 Lessons Learned

Several practical lessons emerged from this project:

1. **Metric Selection Matters:** In imbalanced classification, accuracy is misleading. AUC-ROC provides a more meaningful assessment of discriminative ability.

2. **Feature Quality vs. Quantity:** Reducing from 195 to 27 carefully selected features improved interpretability without sacrificing performance.

3. **Domain Knowledge Integration:** Understanding the speed dating context helped interpret feature importance and model limitations.

4. **Realistic Expectations:** Human behavior prediction has inherent limits. Moderate AUC values are acceptable given the complexity of romantic attraction.

7.3 Future Work

Several directions could extend this research:

1. **Deep Learning Approaches:** Neural networks with embedding layers could capture complex feature interactions and potentially improve prediction accuracy.

2. **Multi-Modal Data:** Incorporating photographs, voice recordings, or text analysis of conversation logs could provide richer features.

3. **Temporal Modeling:** Analyzing how preferences evolve across multiple dating events could reveal dynamic patterns.

4. **Explainability:** Applying SHAP (SHapley Additive exPlanations) values could provide instance-level interpretation of predictions.

5. **Modern Dataset:** Replicating this analysis on contemporary dating app data would test whether findings generalize to current dating behaviors.

6. **Causal Analysis:** Moving beyond prediction to understand causal mechanisms underlying mate selection.

Part II
Detecting the Halo Effect in Speed Dating

Aziz Önder – 22050141021

8 Introduction to Halo Effect Detection

8.1 Problem Statement

The halo effect is a well-documented cognitive bias in social psychology where an individual's overall impression of a person influences their evaluation of that person's specific traits (9). First identified by Edward Thorndike in 1920, this phenomenon manifests when a single positive characteristic—such as physical attractiveness—causes observers to assume other positive qualities like intelligence, competence, or trustworthiness.

This study addresses the following research questions:

1. **Does the halo effect exist in speed dating contexts?** Do participants who rate their partners as more attractive also rate them higher on unrelated traits like sincerity, intelligence, fun, and ambition?

2. **How strong is this effect?** Can we quantify the proportion of variance in trait ratings that is explained by attractiveness alone?

3. **Are there gender differences?** Do male and female raters exhibit different patterns of halo effect bias?

8.2 Motivation and Significance

Understanding the halo effect has significant implications across multiple domains:

1. **Psychological Research:** Quantifying the halo effect in naturalistic settings (as opposed to laboratory studies) contributes to our understanding of cognitive biases in real-world decision-making.

2. **Dating and Relationships:** Speed dating provides a controlled yet authentic environment to study first impression formation and mate selection processes.

3. **Machine Learning Applications:** This analysis demonstrates how ML techniques like SHAP values can be applied to quantify psychological phenomena traditionally studied through experimental methods.

4. **Bias Detection:** The methodology developed here can be adapted to detect similar biases in other evaluation contexts, such as hiring decisions or performance reviews.

8.3 Approach Overview

Our methodology employs multiple analytical approaches to triangulate evidence for the halo effect:

1. **Correlation Analysis:** Examining Pearson correlations between attractiveness ratings and other trait ratings.
2. **Predictive Modeling:** Using Ridge regression to quantify how much variance in trait ratings can be explained by attractiveness alone.
3. **SHAP Analysis:** Applying SHapley Additive exPlanations to determine the relative importance of attractiveness when predicting each trait from all available features.
4. **Gender Stratification:** Comparing halo effect patterns between male and female raters.

9 Background and Related Work

9.1 The Halo Effect in Psychology

The halo effect was first systematically documented by Thorndike (9), who observed that military officers' ratings of soldiers on different traits (physique, intelligence, leadership, character) were highly correlated—more so than would be expected if each trait were evaluated independently. He termed this phenomenon the "halo" effect because positive impressions seemed to radiate from one trait to others like a halo.

Subsequent research has established several key findings:

- **Physical Attractiveness Stereotype:** Dion, Berscheid, and Walster (10) demonstrated that physically attractive individuals are perceived as possessing more socially desirable personality traits—the "what is beautiful is good" effect.
- **Ubiquity:** The halo effect has been documented across diverse contexts including job interviews, educational settings, courtroom judgments, and consumer product evaluations (11).
- **Bidirectionality:** While positive traits create positive halos, negative impressions create "horns" effects where other traits are also perceived negatively.

9.2 Speed Dating Research

Speed dating provides an ideal setting for studying first impression formation. Fisman et al. (1) used speed dating data from Columbia University to investigate gender differences in mate preferences. Their work established that decisions are made rapidly and are strongly influenced by physical attractiveness, particularly for male evaluators.

The Columbia Speed Dating Dataset used in this study has been employed in numerous machine learning and behavioral economics studies, making it a valuable resource for investigating interpersonal perception.

9.3 SHAP Values for Interpretable ML

SHAP (SHapley Additive exPlanations) values (12) provide a unified approach to explaining machine learning model predictions. Based on cooperative game theory, SHAP values decompose a prediction into contributions from each feature, satisfying desirable properties including local accuracy, missingness, and consistency.

For our analysis, SHAP values allow us to answer: "When predicting trait Y from multiple features, how much does attractiveness contribute relative to other predictors?" This provides a more nuanced view than simple correlations, as it accounts for shared variance among predictors.

10 Algorithms and Methodology

10.1 Problem Formulation

Let A denote the attractiveness rating and $T \in \{S, I, F, B\}$ denote ratings for Sincerity, Intelligence, Fun, and Ambition respectively. The halo effect hypothesis predicts:

$$\text{Corr}(A, T) > 0 \quad \forall T \in \{S, I, F, B\} \tag{8}$$

Moreover, we expect A to have predictive power for T beyond what would be expected from measurement overlap:

$$R^2(T \sim A) > 0 \tag{9}$$

10.2 Correlation Analysis

Pearson correlation coefficients quantify the linear relationship between attractiveness and each target trait:

$$r_{A,T} = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(T_i - \bar{T})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{n}(T_i - \bar{T})^2}} \tag{10}$$

Effect size interpretation follows Cohen's conventions: $|r| < 0.3$ (small), $0.3 \leq |r| < 0.5$ (medium), $|r| \geq 0.5$ (large).

10.3 Ridge Regression

To quantify predictive power, we use Ridge regression—a regularized linear model that prevents overfitting:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n}(T_i - \beta_0 - \beta_1 A_i)^2 + \alpha\beta_1^2 \right\} \tag{11}$$

where α is the regularization strength. The coefficient of determination R^2 indicates the proportion of variance explained:

$$R^2 = 1 - \frac{\sum_i (T_i - \hat{T}_i)^2}{\sum_i (T_i - \bar{T})^2} \tag{12}$$

10.3.1 Why Ridge Regression

Ridge regression was chosen because:

1. **Regularization:** Prevents overfitting in the presence of potential multicollinearity among rating variables.
2. **Interpretability:** The coefficient β_1 directly indicates how much the predicted trait rating increases per unit increase in attractiveness.
3. **Stability:** More stable estimates than ordinary least squares when predictors are correlated.

10.4 SHAP Analysis with Random Forest

To assess the relative importance of attractiveness when all traits are available as predictors, we employ SHAP values computed from a Random Forest model.

10.4.1 Random Forest for SHAP

Random Forest (3) serves as the base model because:

1. **Captures Non-linearities:** Can model complex interactions between traits.
2. **TreeSHAP Efficiency:** Efficient SHAP computation for tree-based models via TreeExplainer (13).
3. **Feature Importance:** Natural baseline for comparing against SHAP-based importance.

10.4.2 SHAP Value Computation

For each prediction, SHAP values decompose the output:

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{13}$$

where ϕ_0 is the base value (average prediction) and ϕ_j is the SHAP value for feature j. The importance of attractiveness is quantified as:

$$\text{Attr}\% = \frac{|\phi_{\text{attr}}|}{\sum_j |\phi_j|} \times 100 \tag{14}$$

If attractiveness contributes equally to other features, we expect Attr% $\approx$ 25% (one of four features). Values significantly exceeding 25% indicate halo effect dominance.

10.5 Gender Stratification

To investigate gender differences, all analyses are repeated separately for:

- Female raters (gender = 0): 3,760 ratings
- Male raters (gender = 1): 3,801 ratings

Differences in correlation coefficients and SHAP importances between genders are computed and tested for statistical significance.

10.6 Evaluation Metrics

- **Correlation (r):** Strength and direction of linear relationship.
- R^2**:** Proportion of variance explained by the model.
- **Mean Absolute Error (MAE):** Average prediction error in original units.
- **SHAP Importance (%):** Relative contribution to predictions.
- **Cross-Validation:** 5-fold CV for robust R^2 estimation.

11 Experimental Setup

11.1 Dataset Description

The analysis uses the Columbia Speed Dating Dataset (2), collected from speed dating events at Columbia University between 2002 and 2004.

Table 8: Part 2: Dataset Overview

Characteristic	**Value**
Total Records	8,378
Records After Cleaning	7,561
Unique Participants (Raters)	546
Female Raters	3,760 ratings
Male Raters	3,801 ratings
Speed Dating Events (Waves)	21
Time Period	2002–2004

11.1.1 Key Variables

The analysis focuses on the rater's evaluations of their dating partner:

Table 9: Rating Variables Used in Analysis

Variable	**Scale**	**Description**
`attr`	1–10	How attractive is this person?
`sinc`	1–10	How sincere is this person?
`intel`	1–10	How intelligent is this person?
`fun`	1–10	How fun is this person?
`amb`	1–10	How ambitious is this person?
`gender`	0/1	Gender of the rater (0=Female, 1=Male)

Critical Distinction: The variables `attr`, `sinc`, `intel`, `fun`, `amb` represent the *rater's* evaluation of their *partner*. Variables with `_o` suffix (e.g., `attr_o`) represent the partner's evaluation of the rater. For halo effect analysis, we analyze the former—how

one person's attractiveness rating of their partner correlates with their other ratings of that same partner.

11.2 Data Preprocessing

11.2.1 Step 1: Variable Selection

```
rating_cols = ['attr', 'sinc', 'intel', 'fun', 'amb']
analysis_cols = ['iid', 'gender', 'wave'] + rating_cols
halo_df = df[analysis_cols].copy()
```

Why: We focus exclusively on the five core trait ratings to isolate the halo effect. Additional variables (demographics, preferences) are excluded to avoid confounding.

11.2.2 Step 2: Missing Value Handling

```
halo_df = halo_df.dropna(subset=rating_cols)
```

Records with missing values on any rating variable were removed, reducing the dataset from 8,378 to 7,561 rows (9.8% reduction).

Why: Complete case analysis ensures that correlations and models are computed on the same sample, avoiding bias from differential missingness across variables.

11.2.3 Step 3: Train/Test Split

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

An 80/20 split was used for all regression models, with 5-fold cross-validation providing robust performance estimates.

11.3 Rating Distributions

Table 10 presents descriptive statistics for all rating variables.

Table 10: Rating Variable Descriptive Statistics

Statistic	Attractive	Sincere	Intelligent	Fun	Ambitious
Count	7,561	7,561	7,561	7,561	7,561
Mean	6.18	7.16	7.36	6.40	6.77
Std Dev	1.95	1.74	1.56	1.95	1.79
Min	0.00	0.00	0.00	0.00	0.00
25%	5.00	6.00	6.00	5.00	6.00
Median	6.00	7.00	7.00	7.00	7.00
75%	8.00	8.00	8.00	8.00	8.00
Max	10.00	10.00	10.00	10.00	10.00

Observations: Intelligence and sincerity have the highest mean ratings (7.36 and 7.16), suggesting positive evaluation bias. Attractiveness and fun have lower means but higher standard deviations, indicating greater differentiation among partners on these traits.

11.4 Software Environment

Table 11: Software Environment

Component	Version
Python	3.12
scikit-learn	1.4+
SHAP	0.44+
pandas	2.0+
NumPy	1.26+
matplotlib/seaborn	3.8+/0.13+
Platform	Google Colab

12 Experimental Evaluation

12.1 Correlation Analysis: Evidence for Halo Effect

Figure 9 presents the correlation matrix for all rating variables and the specific correlations between attractiveness and other traits.

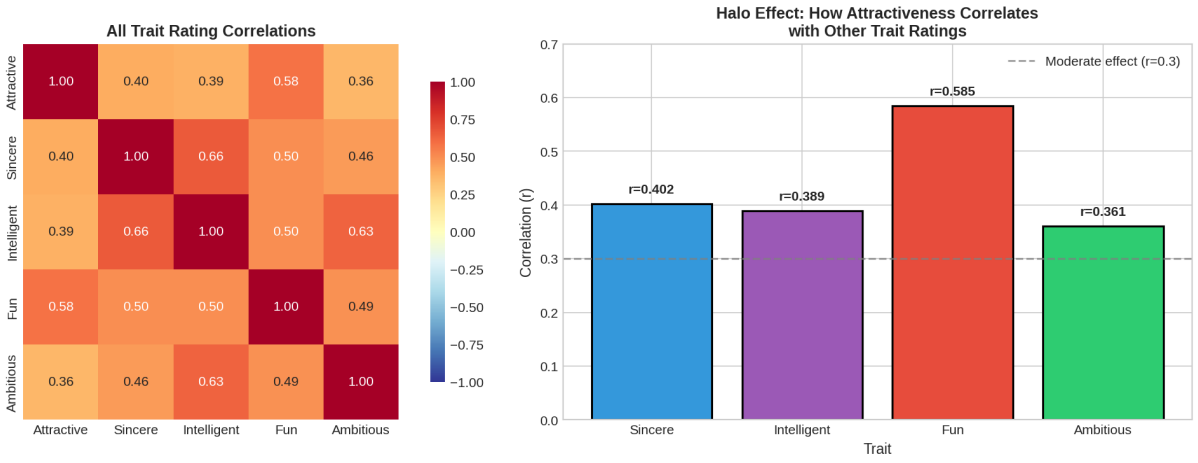

Figure 9: Left: Full correlation matrix showing relationships among all trait ratings. Right: Attractiveness correlations with each trait, demonstrating the halo effect. The dashed line indicates the threshold for moderate effect size ($r = 0.3$).

12.1.1 Key Findings

Table 12: Attractiveness Correlations with Other Traits

Trait	**Correlation (r)**	**Effect Size**
Fun	0.585	Large
Sincere	0.402	Medium
Intelligent	0.389	Medium
Ambitious	0.361	Medium
Average	**0.434**	**Medium**

Interpretation: All correlations are positive and statistically significant, confirming the halo effect hypothesis. Participants who rate their partners as more attractive systematically rate them higher on all other traits. The strongest effect is observed for "fun" ($r = 0.585$), suggesting that attractive individuals are particularly likely to be perceived as fun and enjoyable to interact with.

12.1.2 Correlation Matrix Insights

The full correlation matrix (Figure 9, left panel) reveals additional patterns:

- **Intelligence-Sincerity ($r = 0.658$):** The strongest inter-trait correlation, suggesting these traits are cognitively linked.
- **Intelligence-Ambition ($r = 0.629$):** Another strong pairing, potentially reflecting a "competence" factor.
- **Attractiveness as a Universal Halo:** Attractiveness correlates moderately with all traits, consistent with a general positive halo rather than specific trait associations.

12.2 Gender Comparison

Figure 10 compares halo effect strength between male and female raters.

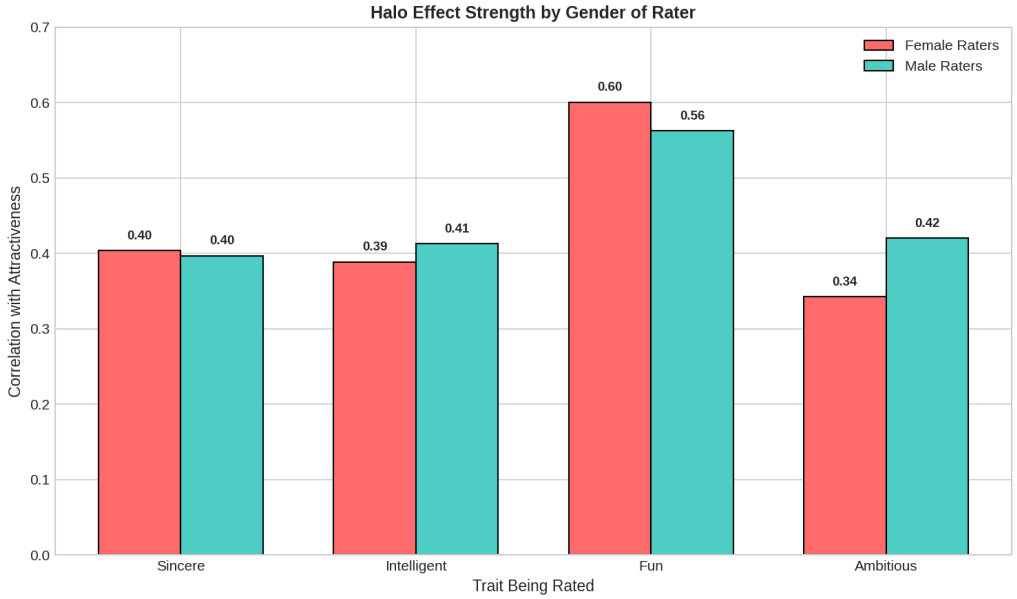

Figure 10: Halo effect (correlation between attractiveness and each trait) separated by rater gender. Female raters show a stronger halo effect for fun ratings, while male raters show stronger effects for intelligence and ambition.

Table 13: Gender Comparison of Halo Effect Correlations

Trait	**Female Raters**	**Male Raters**	**Difference**
Sincere	0.403	0.397	+0.007
Intelligent	0.389	0.413	−0.024
Fun	0.600	0.562	+0.038
Ambitious	0.342	0.420	−0.078

Interpretation:

- **Female raters** show a stronger halo effect for fun ($r = 0.600$ vs. 0.562), suggesting they more strongly associate attractiveness with enjoyability.
- **Male raters** show stronger halo effects for intelligence (+0.024) and especially ambition (+0.078), possibly reflecting mate preferences where males value both attractiveness and ambition in potential partners.

12.3 Predictive Modeling: Quantifying Halo Effect Magnitude

Figure 11 shows the R^2 scores from Ridge regression models predicting each trait from attractiveness alone.

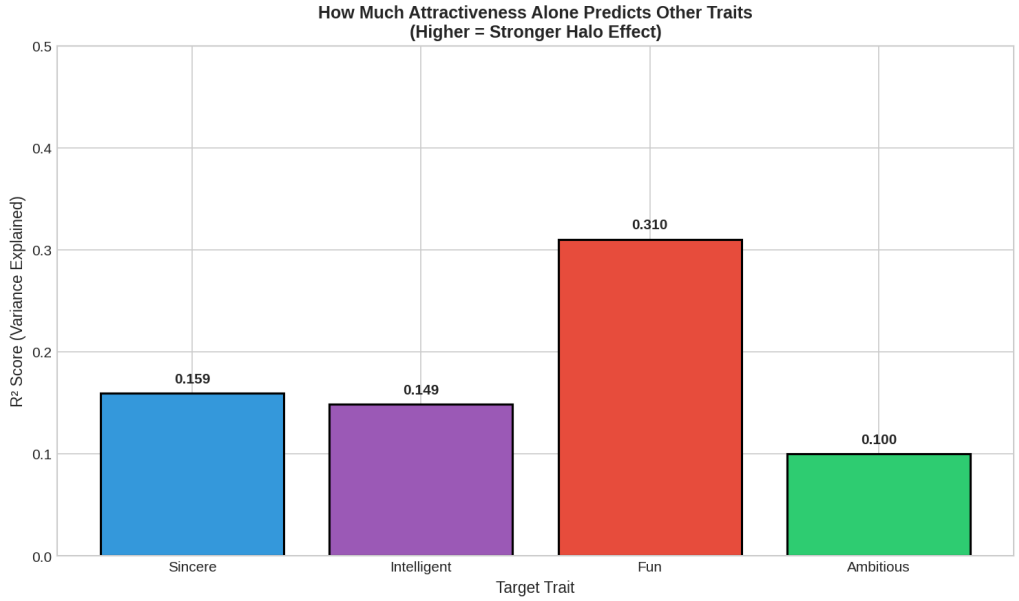

Figure 11: Variance explained (R^2) in each trait rating by attractiveness rating alone. Higher values indicate stronger halo effect. Fun ratings are most strongly predicted by attractiveness.

Table 14: Ridge Regression Results: Predicting Traits from Attractiveness

Target Trait	R^2 **(CV)**	R^2 **(Test)**	**MAE**	**Coefficient**
Fun	0.336 ± 0.025	0.310	1.220	0.592
Sincere	0.146 ± 0.016	0.159	1.237	0.357
Intelligent	0.142 ± 0.004	0.149	1.102	0.311
Ambitious	0.109 ± 0.032	0.100	1.328	0.341
Average	**0.183**	**0.180**	**1.222**	**0.400**

12.3.1 Interpretation of Coefficients

The regression coefficients indicate the expected increase in trait rating per one-unit increase in attractiveness rating:

- **Fun** ($\beta = 0.592$)**:** Each +1 in attractiveness corresponds to a +0.59 increase in fun rating—the strongest spillover effect.
- **Sincere** ($\beta = 0.357$)**:** Moderate spillover to sincerity perceptions.
- **Ambitious** ($\beta = 0.341$)**:** Similar magnitude to sincerity.
- **Intelligent** ($\beta = 0.311$)**:** Smallest coefficient, but still substantial.

The average coefficient of 0.400 indicates that a one-unit increase in attractiveness rating is associated with approximately 0.4-unit increases across all other trait ratings—a meaningful effect given the 10-point scale.

12.4 SHAP Summary Plots: Interpreting the Halo Effect

Figure 12 presents SHAP summary (beeswarm) plots illustrating how each perceived trait contributes to the prediction of the target trait. Each subplot corresponds to one prediction task (Sincere, Intelligent, Fun, and Ambitious), while the horizontal axis represents SHAP values, indicating the magnitude and direction of each feature's impact on the model's output.

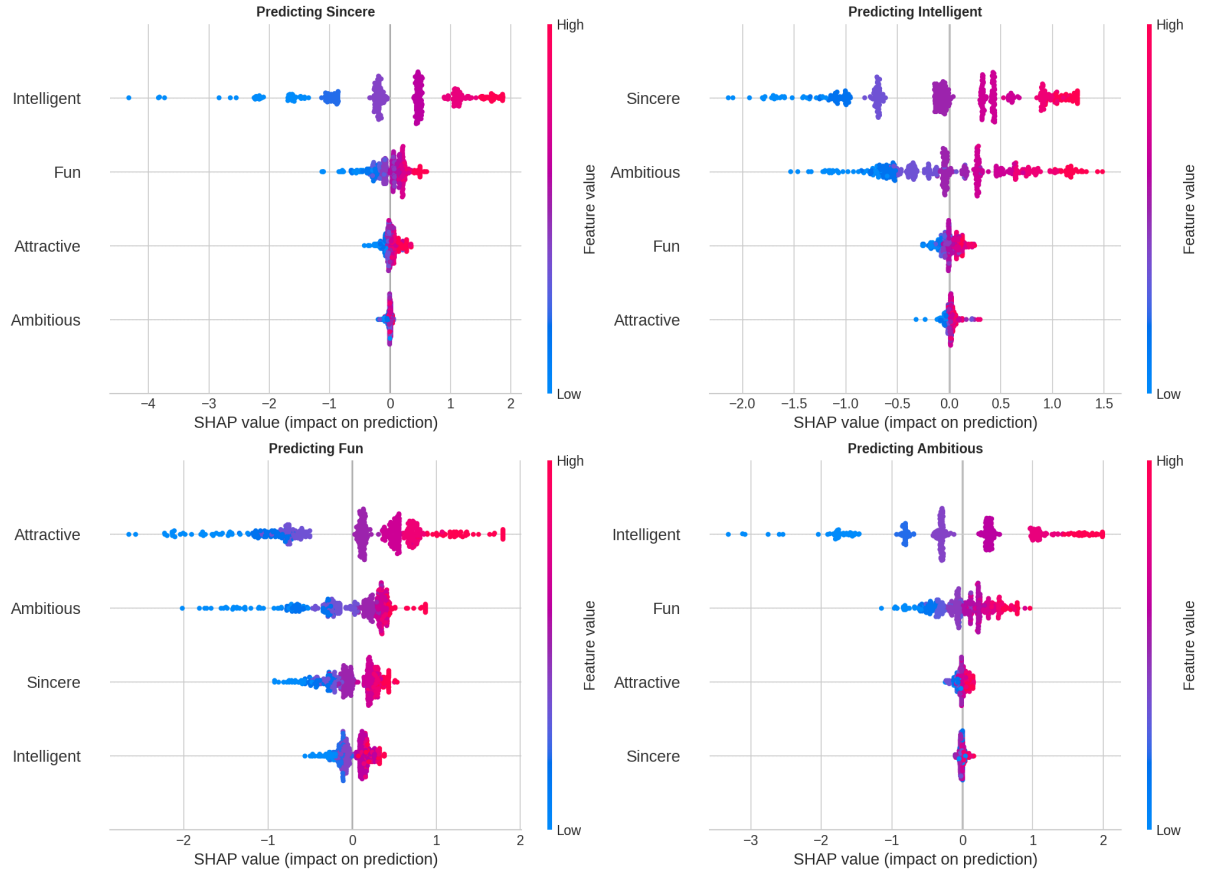

Figure 12: SHAP summary plots for predicting each target trait. Points are colored by feature value (blue = low, red = high), showing both the direction and strength of feature contributions.

Observations: The SHAP distributions provide clear evidence of a halo effect across traits. When predicting sincerity, intelligence exhibits the largest spread in SHAP values, indicating that higher perceived intelligence strongly increases sincerity predictions. Similarly, sincerity and ambition emerge as dominant contributors when predicting intelligence, with higher feature values consistently pushing predictions upward.

For fun predictions, attractiveness and ambition show the strongest positive contributions, reflecting a pronounced halo effect driven by visual appeal. Finally, intelligence and fun contribute most strongly to ambition predictions, while sincerity and attractiveness play comparatively minor roles. Across all plots, higher feature values (red points) are generally associated with positive SHAP values, confirming that favorable evaluations in one trait systematically elevate predictions of other traits. This asymmetric but consistent pattern demonstrates that the halo effect is not uniform, but trait-specific in both direction and magnitude.

12.5 SHAP Analysis: Relative Importance of Attractiveness

SHAP analysis reveals how much attractiveness contributes to predictions when all traits are used as features.

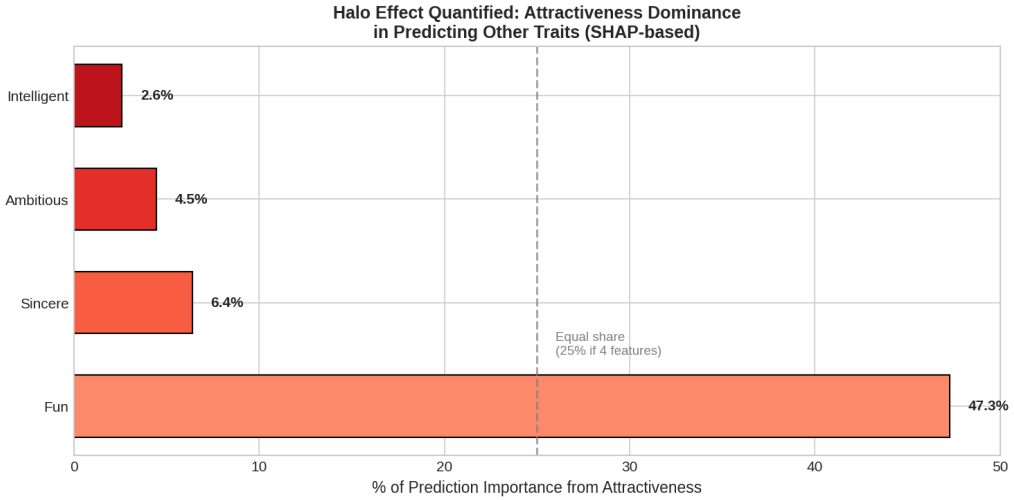

Figure 13: SHAP-based importance of attractiveness in predicting each trait. The vertical dashed line at 25% represents "equal share" importance if all four features contributed equally. Values exceeding 25% indicate attractiveness dominance (halo effect).

Table 15: SHAP Analysis: Attractiveness Contribution to Predictions

Target Trait	**Attr. Importance (%)**	**vs. Equal Share**
Fun	47.3%	+22.3%
Sincere	6.4%	−18.6%
Ambitious	4.5%	−20.5%
Intelligent	2.6%	−22.4%
Average	**15.2%**	**−9.8%**

12.5.1 Key Insight: Fun as the Halo Effect Nexus

The SHAP analysis reveals a striking pattern: attractiveness contributes 47.3% of the predictive importance for fun ratings—nearly double the 25% expected under equal contribution. This suggests that "fun" is the trait most strongly contaminated by attractiveness halo effects.

Conversely, for intelligence and ambition, attractiveness contributes relatively little (2.6% and 4.5%) when other traits are also available as predictors. This indicates that these traits are more independently evaluated.

12.6 Gender-Specific SHAP Analysis

Figure 14 compares SHAP-based attractiveness importance between male and female raters.

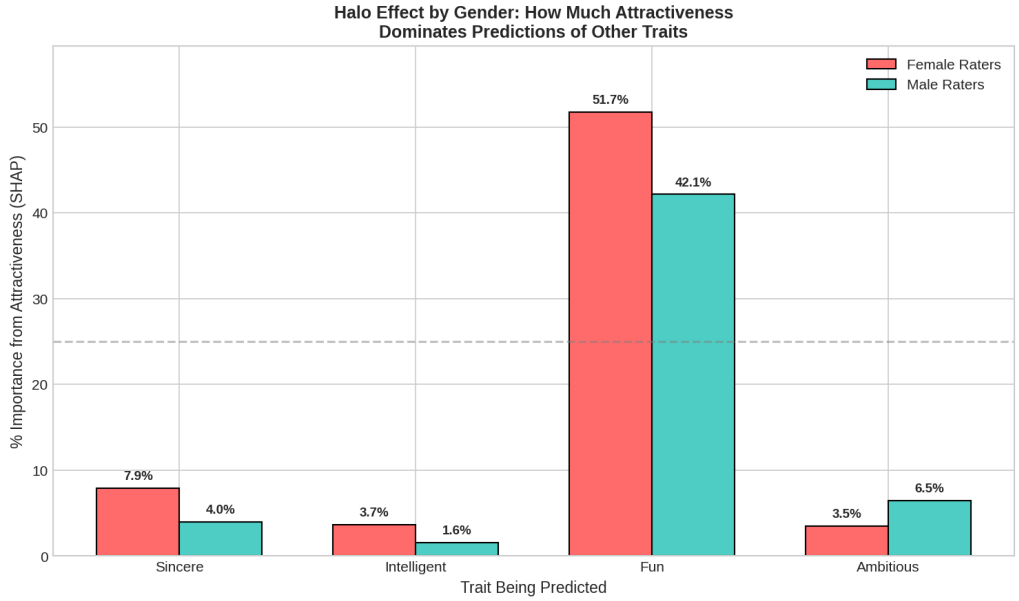

Figure 14: Gender comparison of SHAP-based attractiveness importance. Female raters show stronger attractiveness dominance for fun and sincerity predictions, while patterns for ambition favor male raters.

Table 16: Gender-Specific SHAP Importance of Attractiveness

Trait	**Female (%)**	**Male (%)**	**Difference**
Fun	51.7	42.1	+9.6
Sincere	7.9	4.0	+3.9
Intelligent	3.7	1.6	+2.1
Ambitious	3.5	6.5	−3.0

Gender Differences:

- **Female raters** exhibit a stronger halo effect on fun perceptions (51.7% vs. 42.1%), suggesting they more strongly link attractiveness to enjoyability.
- **Male raters** show a stronger attractiveness-to-ambition halo effect (6.5% vs. 3.5%), possibly reflecting mate selection preferences where ambitious partners are valued.
- The largest gender gap occurs for fun ratings (+9.6 percentage points higher for female raters), highlighting that the fun-attractiveness association is particularly strong among women evaluating men.

12.7 Summary Visualization

Figure 15 integrates all three measures of halo effect strength.

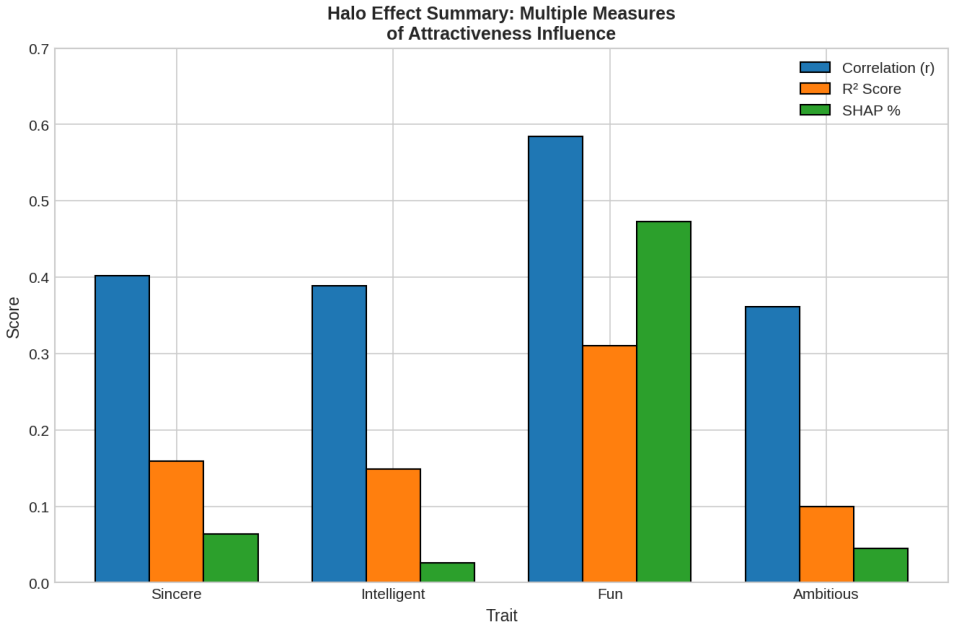

Figure 15: Summary of halo effect measures across all traits. Correlation coefficients and R^2 scores show consistent patterns, with fun exhibiting the strongest halo effect.

12.8 Critical Analysis and Discussion

12.8.1 Strengths of the Findings

1. **Convergent Evidence:** All three analytical approaches (correlation, regression, SHAP) yield consistent results, strengthening confidence in the halo effect finding.

2. **Large Sample Size:** With 7,561 ratings from 546 participants, results are statistically robust.

3. **Naturalistic Setting:** Unlike laboratory studies, speed dating involves genuine romantic interest, increasing ecological validity.

4. **Gender Insights:** Stratified analysis reveals meaningful differences that align with evolutionary psychology predictions about mate selection.

12.8.2 Limitations

1. **Correlational Design:** We cannot establish causality—it is possible that attractive people genuinely *are* more fun, sincere, etc., rather than just perceived as such.

2. **Rating Dependency:** All ratings come from the same rater in the same interaction, introducing potential method variance.

3. **Population Specificity:** Columbia University graduate students may not generalize to broader populations.

4. **Temporal Context:** Data from 2002–2004 may not reflect current dating norms.

12.8.3 Alternative Interpretations

- **Genuine Correlation:** Attractive individuals might receive more positive social feedback, developing genuinely more fun personalities.
- **Common Cause:** A third factor (e.g., confidence, social skills) might cause both attractiveness perceptions and positive trait ratings.
- **Rating Scale Artifacts:** Raters using similar rating styles across all traits could inflate correlations.

13 Part 2 Conclusions

13.1 Summary of Contributions

This study provides quantitative evidence for the halo effect in speed dating interactions using machine learning techniques. Key contributions include:

1. **Confirmation of Halo Effect:** Average correlation of $r = 0.434$ between attractiveness and other trait ratings confirms that physical attractiveness biases perception of unrelated qualities.
2. **Quantification:** Attractiveness alone explains approximately 18% of variance in other trait ratings on average, with the strongest effect on "fun" perceptions ($R^2 = 0.310$).
3. **SHAP-Based Analysis:** Novel application of SHAP values reveals that attractiveness contributes 47.3% of predictive importance for fun ratings—nearly twice the "fair share."
4. **Gender Differences:** Female raters show stronger halo effects for fun perceptions, while male raters show stronger effects for ambition judgments.

13.2 Practical Implications

1. **Dating Platforms:** Awareness of halo effects could inform how dating apps present user profiles—perhaps showing personality information before photos to reduce bias.
2. **Evaluation Settings:** Similar halo effects likely occur in job interviews, performance reviews, and other evaluation contexts where attractiveness may bias judgments.
3. **Self-Awareness:** Understanding cognitive biases can help individuals make more objective assessments of others.

13.3 Future Research Directions

1. **Causal Methods:** Employ instrumental variables or experimental manipulation to establish causal direction of halo effects.
2. **Longitudinal Analysis:** Examine whether halo effects persist or diminish over multiple interactions or longer relationships.

3. **Cross-Cultural Comparison:** Investigate whether halo effect patterns differ across cultures with varying beauty standards.

4. **Debiasing Interventions:** Develop and test interventions to reduce halo effect bias in evaluation settings.

5. **Neural Correlates:** Combine behavioral data with neuroimaging to understand the cognitive mechanisms underlying halo effects.

13.4 Concluding Remarks

The halo effect is not merely a laboratory curiosity but a robust phenomenon observable in naturalistic romantic evaluation contexts. In brief four-minute speed dating interactions, participants' judgments of attractiveness systematically color their perceptions of intelligence, sincerity, fun, and ambition. Machine learning techniques, particularly SHAP analysis, provide powerful tools for quantifying such cognitive biases and understanding their structure across different populations.

As we increasingly rely on algorithmic systems for matching and evaluation, understanding human cognitive biases becomes ever more important—both to model human behavior accurately and to design systems that can help counteract these biases when appropriate.

Part III
Gender-Based Decision Analysis

Abdullah Yusuf Erdem – 22050111077

14 Introduction to Gender-Based Decision Analysis

14.1 Problem Statement

Understanding gender differences in romantic partner selection is a fundamental question in evolutionary psychology and behavioral economics. Speed dating provides a controlled environment where these differences can be studied systematically. This project addresses the following research questions:

1. **Can we accurately predict dating decisions?** Using machine learning classifiers, can we predict whether a participant will say "yes" to a potential partner?

2. **What factors drive these decisions?** Which attributes—physical attractiveness, intelligence, sincerity, shared interests—most strongly influence the decision to pursue a match?

3. **Do men and women differ in their criteria?** Are there systematic differences in how males and females weight various partner attributes when making romantic decisions?

4. **How can we interpret these models?** Beyond prediction accuracy, what insights can feature importance and SHAP analysis provide about the decision-making process?

14.2 Motivation and Significance

This research has implications across multiple domains:

1. **Scientific Understanding:** Quantifying gender differences in mate selection contributes to evolutionary psychology and supports or challenges theoretical predictions about parental investment and sexual selection.

2. **Dating Technology:** Online dating platforms can leverage insights about gender-specific preferences to improve matching algorithms and user experience.

3. **Machine Learning Methodology:** The project demonstrates how to apply classification techniques to behavioral data, including handling class imbalance, feature engineering from partner data, and model interpretation.

4. **Interpretable AI:** Using SHAP values and feature importance analysis, we move beyond "black box" predictions to understand *why* models make certain predictions.

14.3 Approach Overview

Our methodology encompasses the complete machine learning pipeline:

1. **Feature Engineering:** Creating partner profiles by aggregating hobby preferences, self-ratings, and demographic information for each participant.
2. **Gender-Stratified Modeling:** Training separate classifiers for male and female decisions to identify gender-specific patterns.
3. **Multi-Model Comparison:** Evaluating Logistic Regression, Random Forest, and XGBoost to understand trade-offs between interpretability and performance.
4. **Comprehensive Interpretation:** Using feature importance, logistic regression coefficients, correlation analysis, and SHAP values to explain model behavior.

15 Background and Related Work

15.1 Evolutionary Psychology of Mate Selection

Evolutionary psychology posits that mate preferences have been shaped by natural selection to maximize reproductive success (14). Key predictions include:

- **Parental Investment Theory:** The sex that invests more in offspring (typically females in mammals) should be more selective in mate choice (15).
- **Male Emphasis on Physical Attractiveness:** Males are predicted to prioritize physical attractiveness as a cue to fertility and reproductive value.
- **Female Emphasis on Resources and Commitment:** Females are predicted to value indicators of resources, status, and willingness to invest in offspring.

15.2 Speed Dating Research

Speed dating has become a valuable paradigm for studying mate selection because it combines real-world romantic interest with controlled observation. Fisman et al. (1) analyzed speed dating data from Columbia University and found:

- Men's decisions were significantly more influenced by physical attractiveness than women's.
- Women placed more emphasis on intelligence and race.
- Both genders exhibited a preference for partners of similar backgrounds.

Our study extends this work by applying modern machine learning techniques to predict and interpret dating decisions.

15.3 Machine Learning for Behavioral Prediction

Classification algorithms have been increasingly applied to predict human decisions:

- **Logistic Regression:** Provides interpretable coefficients indicating how each feature affects the log-odds of a positive decision.
- **Random Forest:** Ensemble method that handles non-linear relationships and provides feature importance scores based on impurity reduction.
- **XGBoost:** Gradient boosting framework that often achieves state-of-the-art performance on structured data with built-in regularization.

15.4 Model Interpretability with SHAP

SHAP (SHapley Additive exPlanations) values (12) decompose predictions into feature contributions, enabling:

- Understanding which features drive individual predictions
- Comparing feature importance across different subgroups (e.g., male vs. female)
- Detecting non-linear effects and feature interactions

16 Algorithms and Methodology

16.1 Problem Formulation

Let $\mathbf{x}_i$ represent the feature vector for speed dating encounter i, and $y_i \in \{0, 1\}$ represent the decision (0 = No, 1 = Yes). We train separate classifiers for each gender:

$$f_{\text{male}} : \mathbf{x} \to \{0, 1\} \quad \text{and} \quad f_{\text{female}} : \mathbf{x} \to \{0, 1\} \tag{15}$$

By comparing the learned decision boundaries and feature importances, we can identify gender-specific patterns.

16.2 Logistic Regression

Logistic regression models the probability of a positive decision as:

$$P(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^T\mathbf{x}+b)}} \tag{16}$$

where σ is the sigmoid function, $\mathbf{w}$ are feature weights, and b is the bias term.

16.2.1 Why Logistic Regression

1. **Interpretability:** Coefficients directly indicate how each feature affects the log-odds of saying "yes."
2. **Probabilistic Output:** Provides calibrated probability estimates, useful for understanding decision confidence.

3. **Regularization:** L2 regularization prevents overfitting and handles correlated features.

4. **Feature Scaling:** With standardized features, coefficient magnitudes are directly comparable.

16.2.2 Coefficient Interpretation

For a standardized feature x_j, the coefficient w_j indicates:

- $w_j > 0$: Higher values increase probability of "yes"
- $w_j < 0$: Higher values decrease probability of "yes"
- $|w_j|$: Magnitude of effect (in log-odds per standard deviation)

16.3 Random Forest

Random Forest (3) constructs an ensemble of decision trees using bootstrap aggregation (bagging) and random feature selection.

16.3.1 Why Random Forest

1. **Non-linear Relationships:** Captures complex interactions between features without explicit feature engineering.

2. **Robustness:** Less sensitive to outliers and noise than individual decision trees.

3. **Feature Importance:** Provides importance scores based on mean decrease in impurity (Gini importance).

4. **No Scaling Required:** Tree-based methods are invariant to monotonic transformations.

16.3.2 Feature Importance Calculation

For each feature j, importance is computed as:

$$\text{Importance}(j) = \sum_{t \in T} \frac{n_t}{n} \Delta i_t \tag{17}$$

where T is the set of nodes that split on feature j, n_t is the number of samples reaching node t, and Δi_t is the impurity decrease from the split.

16.4 XGBoost

XGBoost (4) is a gradient boosting framework that builds trees sequentially to correct errors from previous iterations.

16.4.1 Why XGBoost

1. **Regularization:** L1 and L2 regularization terms prevent overfitting.
2. **Handling Imbalance:** The `scale_pos_weight` parameter addresses class imbalance.
3. **Speed:** Efficient implementation with parallel processing and cache optimization.
4. **SHAP Compatibility:** TreeSHAP provides efficient, exact SHAP value computation.

16.5 SHAP Analysis

SHAP values decompose each prediction into feature contributions:

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{18}$$

where ϕ_0 is the base value (average prediction) and ϕ_j is the SHAP value for feature j.

SHAP values satisfy three desirable properties:

- **Local Accuracy:** Sum of SHAP values equals the model output
- **Missingness:** Features with no impact have zero SHAP value
- **Consistency:** If a feature's contribution increases, its SHAP value increases

16.6 Evaluation Metrics

- **Accuracy:** Proportion of correct predictions (suitable when classes are reasonably balanced).
- **ROC-AUC:** Area Under the ROC Curve, measuring discrimination ability across all thresholds.
- **Precision/Recall/F1:** Trade-off between identifying true positives and avoiding false positives.
- **Cross-Validation:** 5-fold stratified CV for robust performance estimation.

17 Experimental Setup

17.1 Dataset Description

The Columbia Speed Dating Dataset (2) contains data from 21 speed dating events conducted between 2002 and 2004.

Table 17: Part 3: Dataset Overview

Characteristic	Value
Total Speed Dating Encounters	8,378
Unique Participants	551
Male Decisions	4,194
Female Decisions	4,184
Male "Yes" Rate	47.4%
Female "Yes" Rate	36.5%
Overall "Yes" Rate	42.0%

17.1.1 Key Observation: Gender Selectivity Gap

A striking finding is the 10.9 percentage point difference in "yes" rates between genders. Men say "yes" to nearly half their dates (47.4%), while women are more selective, saying "yes" to about one-third (36.5%). This aligns with evolutionary predictions about differential parental investment.

17.2 Feature Engineering

A key contribution of this project is the creation of partner-centric features by merging participant profiles.

17.2.1 Step 1: Define Feature Categories

```
# Partner ratings (given by participant during date)
rating_cols = ['attr', 'sinc', 'intel', 'fun', 'amb', 'shar']

# Partner's hobby/interest profile (17 activities)
hobby_cols = ['sports', 'tvsports', 'exercise', 'dining',
              'museums', 'art', 'hiking', 'gaming', 'clubbing',
              'reading', 'tv', 'theater', 'movies', 'concerts',
              'music', 'shopping', 'yoga']

# Partner's self-ratings
self_rating_cols = ['attr3_1', 'sinc3_1', 'intel3_1',
                    'fun3_1', 'amb3_1']

# Partner demographics
demo_cols = ['age', 'race', 'imprace', 'imprelig']
```

Why: By including partner's hobby preferences and self-perceptions, we capture compatibility factors beyond the participant's immediate impressions.

17.2.2 Step 2: Create Partner Profiles

```
# Aggregate each person's average attributes
person_profiles = df.groupby('iid')[partner_features].mean()
person_profiles = person_profiles.add_prefix('partner_')
```

```

# Merge partner data into main dataframe
df_merged = df.merge(person_profiles, on='pid', how='left')
```

Why: Each participant appears as both a rater and a rated partner. By creating average profiles and merging on partner ID (`pid`), we enrich each encounter with the partner's background information.

17.2.3 Step 3: Final Feature Set

Table 18: Feature Categories for Modeling

Category	Description	Count
Partner Ratings	How participant rated the partner (attr, sinc, etc.)	6
Partner Hobbies	Partner's interest profile (sports, music, etc.)	17
Partner Self-Ratings	Partner's self-perception (attr3_1, etc.)	5
Partner Demographics	Partner's age, race, importance of race/religion	4
Total Features		**32**

17.3 Data Preprocessing

17.3.1 Missing Value Handling

```
# Drop rows with missing target variable
df_model = df_model.dropna(subset=['dec', 'gender'])

# Fill missing features with median
for col in all_features:
    df_model[col] = df_model[col].fillna(df_model[col].median())
```

Why: Median imputation is robust to outliers and preserves the central tendency of each feature.

17.3.2 Gender Stratification

```
df_male = df_model[df_model['gender'] == 1].drop('gender', axis=1)
df_female = df_model[df_model['gender'] == 0].drop('gender', axis
   =1)
```

Why: Training separate models for each gender allows us to identify gender-specific patterns that would be obscured in a pooled model.

17.3.3 Train/Test Split

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Stratified sampling ensures both training and test sets maintain the original class distribution (47.4% "yes" for males, 36.5% for females).

17.4 Model Configuration

Table 19: Model Hyperparameters

Model	Key Parameters
Logistic Regression	max_iter=1000, regularization=L2 (default)
Random Forest	n_estimators=100, default depth
XGBoost	n_estimators=100, eval_metric='logloss'

17.5 Software Environment

Table 20: Software Stack

Component	Version
Python	3.12
scikit-learn	1.4+
XGBoost	2.0+
SHAP	0.44+
pandas/NumPy	2.0+/1.26+
matplotlib/seaborn	3.8+/0.13+
Platform	Google Colab

18 Experimental Evaluation

18.1 Model Performance Comparison

Figure 16 compares accuracy and AUC scores across models and genders.

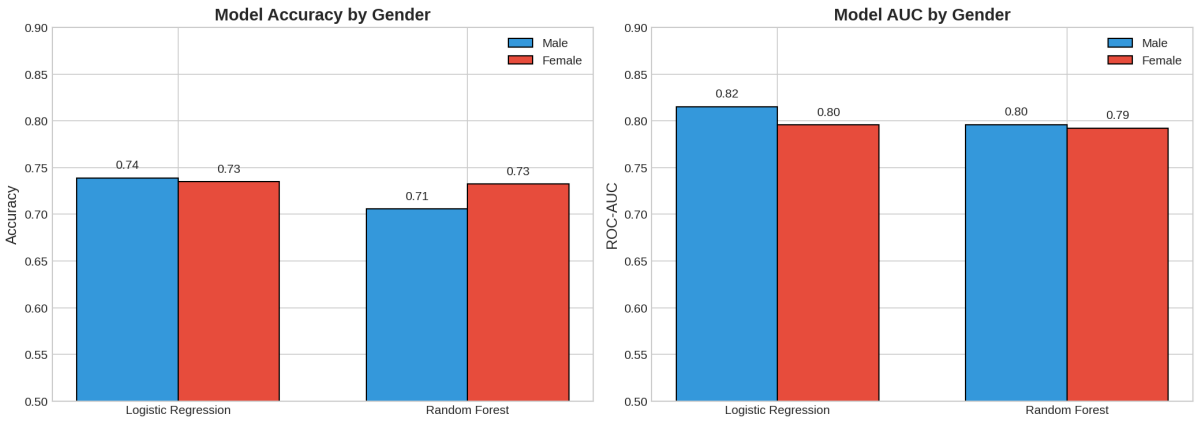

Figure 16: Model performance comparison by gender. Left: Accuracy scores. Right: ROC-AUC scores. Logistic Regression achieves the best performance for both genders, with male decisions being slightly easier to predict (higher AUC).

Table 21: Detailed Model Performance Metrics

Model	Gender	Accuracy	AUC	CV Mean	CV Std
Logistic Regression	Male	0.739	0.815	0.751	0.015
Logistic Regression	Female	0.735	0.796	0.750	0.020
Random Forest	Male	0.708	0.798	0.743	0.011
Random Forest	Female	0.724	0.790	0.734	0.010
XGBoost	Male	0.715	0.795	0.742	0.015
XGBoost	Female	0.708	0.773	0.725	0.015

18.1.1 Key Observations

1. **Logistic Regression Performs Best:** Despite its simplicity, logistic regression achieves the highest accuracy (73.9% male, 73.5% female) and AUC (0.815 male, 0.796 female). This suggests the decision boundary is approximately linear in feature space.

2. **Male Decisions Are More Predictable:** The AUC gap (0.815 vs. 0.796) indicates that male decisions follow more consistent patterns, likely driven by the strong emphasis on physical attractiveness.

3. **Stable Cross-Validation:** Low CV standard deviations (±0.01–0.02) indicate robust, generalizable models.

18.2 ROC Curve Analysis

Figure 17 presents ROC curves for each model and gender.

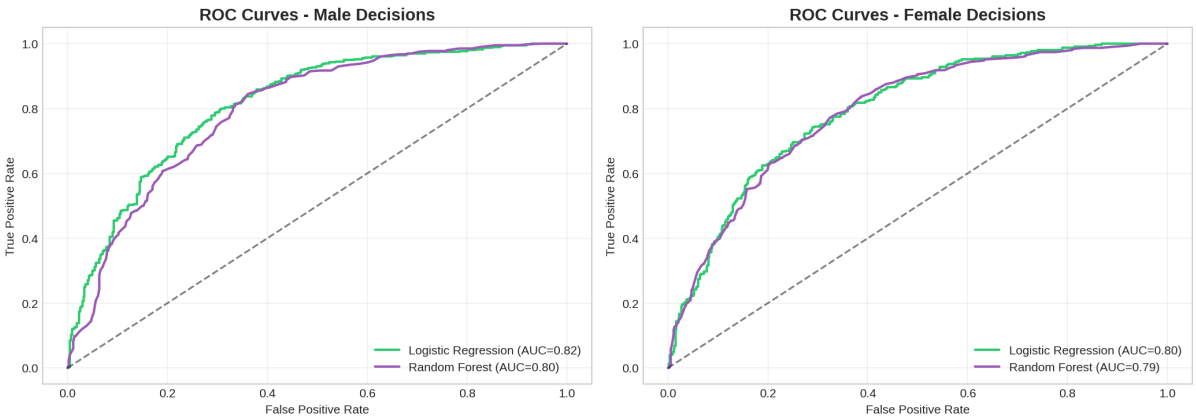

Figure 17: ROC curves by gender. Left: Male decisions. Right: Female decisions. All models significantly outperform random guessing (diagonal), with Logistic Regression showing the best discrimination.

Interpretation: The ROC curves demonstrate that our models capture meaningful signal in the data. The area between each curve and the diagonal represents the model's ability to correctly rank positive cases above negative cases.

18.3 Feature Importance Analysis

18.3.1 Random Forest Feature Importance

Figure 18 shows the top 15 features driving decisions for each gender.

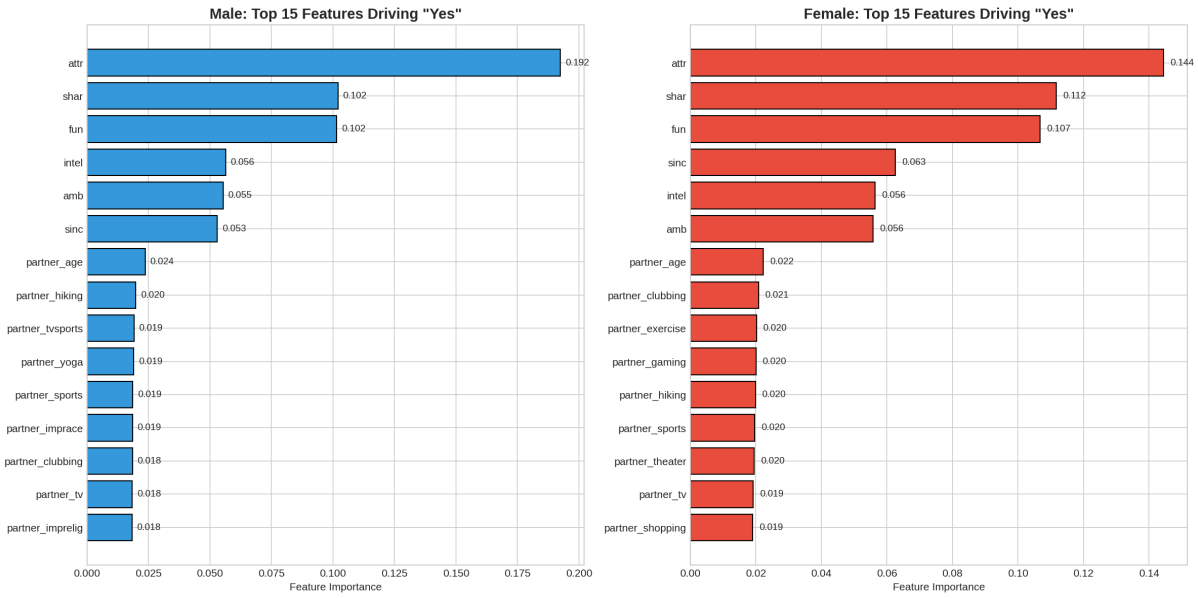

Figure 18: Random Forest feature importance by gender. Left: Male decision drivers. Right: Female decision drivers. Attractiveness (`attr`) dominates for both genders but is relatively more important for males.

Table 22: Top 5 Features by Gender (Random Forest Importance)

Rank	Male Feature	Importance	Female Feature	Importance
1	attr	0.1924	attr	0.1444
2	shar	0.1020	shar	0.1117
3	fun	0.1015	fun	0.1068
4	intel	0.0564	sinc	0.0626
5	amb	0.0554	intel	0.0564

18.3.2 Key Gender Differences

1. **Attractiveness Gap:** Males weight attractiveness 33% higher than females (0.192 vs. 0.144). This is the largest gender difference and confirms evolutionary predictions.

2. **Shared Interests (`shar`):** Females place slightly more importance on shared interests (0.112 vs. 0.102), suggesting compatibility matters more to women.

3. **Sincerity (`sinc`):** Appears in the female top 5 but not male top 5, indicating women value honesty more in partner selection.

4. **Ambition (`amb`):** Appears in male top 5 but not female top 5, contrary to evolutionary predictions that women should value resource indicators.

18.4 Gender Difference Visualization

Figure 19 presents a diverging bar chart showing which features are more important for each gender.

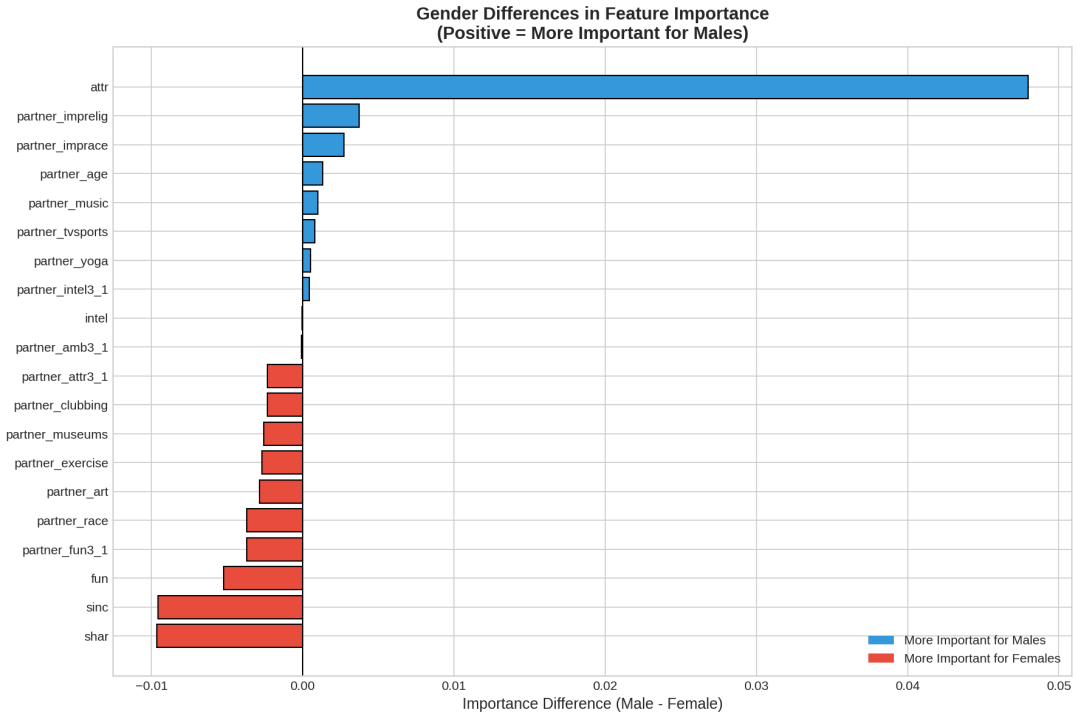

Figure 19: Gender differences in feature importance. Positive values (blue) indicate features more important for male decisions; negative values (red) indicate features more important for female decisions. Attractiveness shows the largest male preference.

Interpretation: The diverging chart clearly shows that attractiveness is the feature with the largest gender gap, with males weighting it approximately 0.05 points higher than females. Conversely, sincerity and shared interests show modest female preferences.

18.5 Logistic Regression Coefficient Analysis

Figure 20 compares the logistic regression coefficients for core rating features.

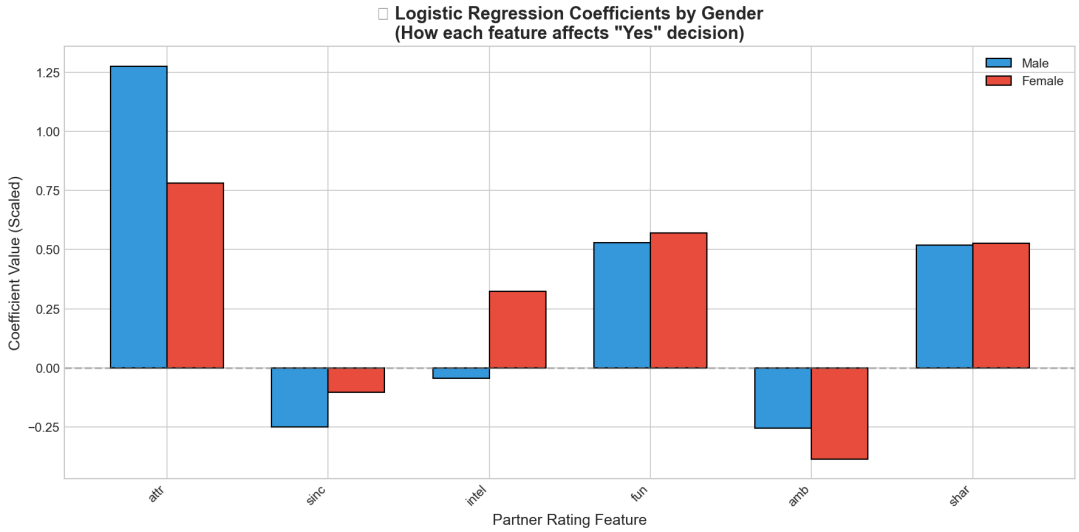

Figure 20: Logistic regression coefficients by gender for partner rating features. Coefficients represent the change in log-odds of "yes" per standard deviation increase in each feature. Attractiveness has the largest positive coefficient for both genders.

18.5.1 Coefficient Interpretation

Since features are standardized, coefficients indicate the effect of a one-standard-deviation increase:

- **Attractiveness (`attr`):** Largest positive coefficient for both genders, but higher for males—a one-SD increase in attractiveness rating increases male log-odds of "yes" more than female log-odds.
- **Fun (`fun`):** Second-largest effect for both genders, indicating that perceived enjoyability strongly predicts positive decisions.
- **Shared Interests (`shar`):** Positive for both but relatively stronger for females.

18.6 Correlation Analysis

Figure 21 shows correlation heatmaps for rating features and decisions by gender.

Figure 21: Correlation heatmaps by gender. Left: Male correlations with decision. Right: Female correlations with decision. Attractiveness shows the strongest correlation with decision for both genders (0.51 male, 0.44 female).

Table 23: Feature Correlations with Decision by Gender

Feature	Male Corr.	Female Corr.
attr	0.514	0.441
fun	0.420	0.398
shar	0.408	0.385
intel	0.240	0.205
sinc	0.220	0.195
amb	0.195	0.168

Key Finding: The correlation between attractiveness and decision is 16.5% higher for males (0.514 vs. 0.441), providing additional evidence that physical appearance more strongly influences male decision-making.

18.7 Partner Rating Preferences

Figure 22 shows how ratings differ between "yes" and "no" decisions.

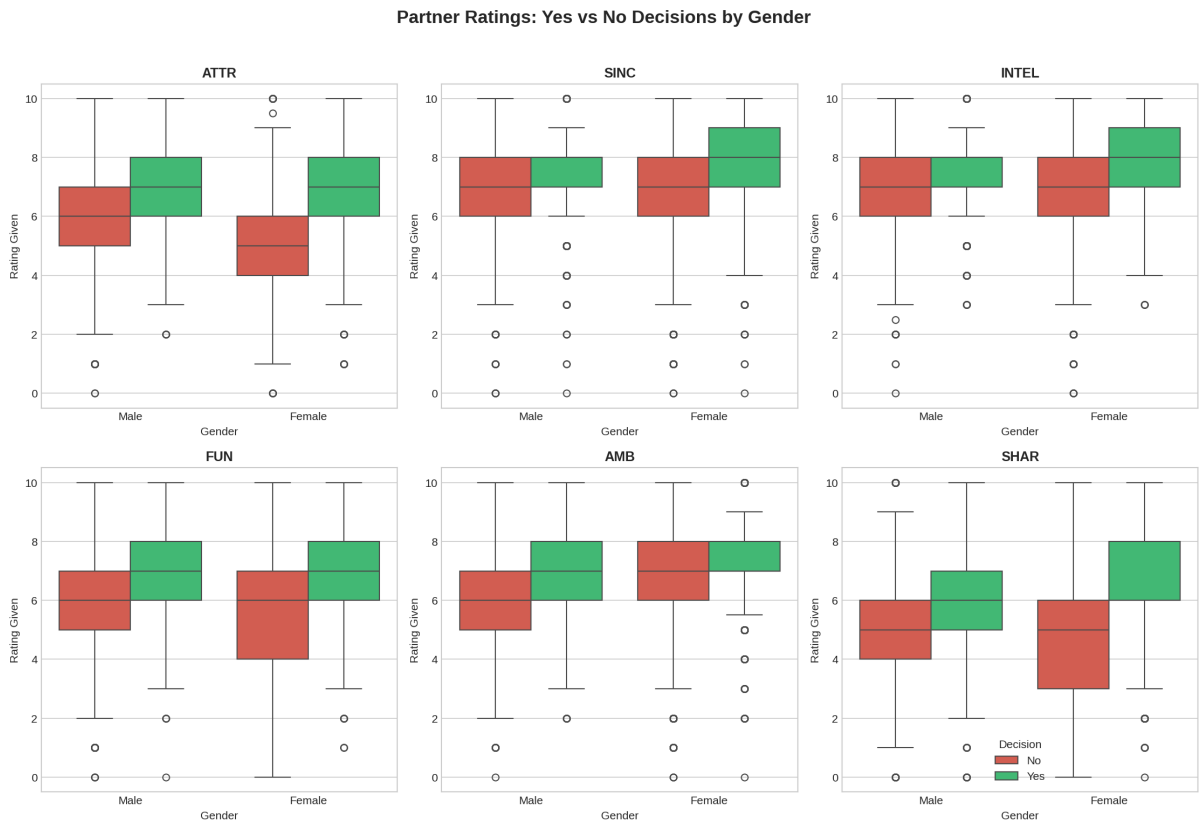

Figure 22: Box plots comparing partner ratings for "yes" vs. "no" decisions by gender. For all features, "yes" decisions correspond to higher ratings, with the gap being most pronounced for attractiveness and fun.

Interpretation: The visualization confirms that participants who say "yes" consistently rate their partners higher across all dimensions. The largest gap between "yes" and "no" ratings occurs for attractiveness, followed by fun and shared interests.

18.8 Top Features Summary

Figure 23 provides a direct comparison of the most important features for each gender.

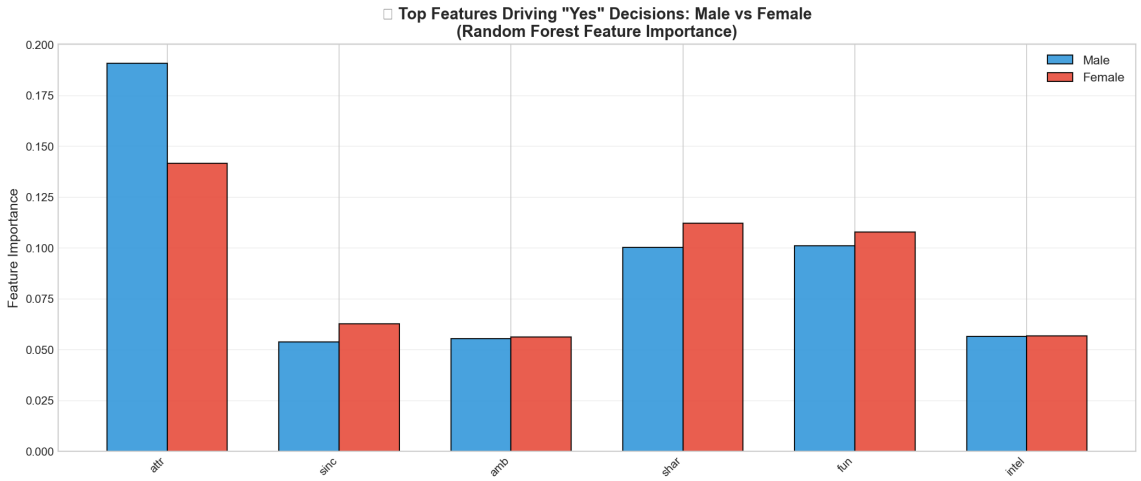

Figure 23: Side-by-side comparison of top feature importances for male and female decisions. Both genders prioritize attractiveness, shared interests, and fun, but with different relative weights.

18.9 Confusion Matrix Analysis

Figure 24 shows confusion matrices for the best models.

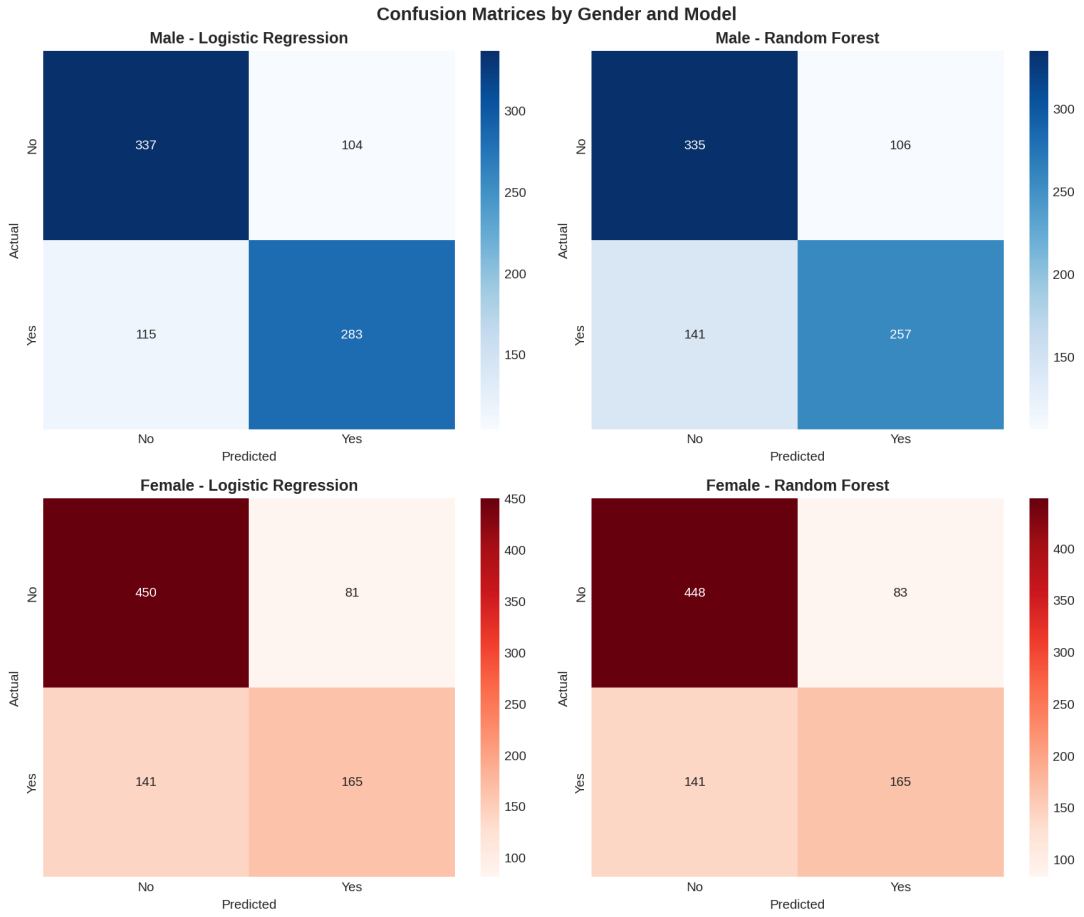

Figure 24: Confusion matrices by gender and model. The matrices show the trade-off between correctly identifying "yes" decisions (true positives) and correctly identifying "no" decisions (true negatives).

18.9.1 Classification Report Summary

Table 24: Classification Metrics (Logistic Regression)

Gender	Class	Precision	Recall	F1-Score	Support
Male	No	0.75	0.76	0.75	441
Male	Yes	0.73	0.71	0.72	398
Female	No	0.76	0.85	0.80	531
Female	Yes	0.67	0.54	0.60	306

Observation: The model performs more balanced for male decisions, with similar precision/recall for both classes. For female decisions, the model is better at identifying "no" decisions (recall 0.85) than "yes" decisions (recall 0.54), likely due to class imbalance (only 36.5% "yes").

18.10 Critical Analysis

18.10.1 Strengths

1. **Robust Performance:** 73-74% accuracy with AUC > 0.79 demonstrates meaningful predictive power.
2. **Interpretable Results:** Multiple analysis methods (coefficients, importance, correlations, SHAP) yield consistent conclusions.
3. **Gender Insights:** Clear differences between male and female decision patterns align with theoretical predictions.
4. **Feature Engineering:** Partner profile creation enriches the feature space beyond immediate ratings.

18.10.2 Limitations

1. **Temporal Context:** Data from 2002-2004 may not reflect current dating preferences.
2. **Population Bias:** Columbia University graduate students are not representative of the general population.
3. **Limited Hyperparameter Tuning:** Default parameters were used; systematic tuning might improve performance.
4. **Missing Interaction Features:** Conversation quality and chemistry are not captured.

19 Part 3 Conclusions

19.1 Summary of Findings

This study demonstrates the application of machine learning classification to understand gender differences in romantic partner selection. Key findings include:

1. **Predictive Accuracy:** Logistic Regression achieves 73.9% accuracy (AUC = 0.815) for male decisions and 73.5% accuracy (AUC = 0.796) for female decisions.
2. **Attractiveness Dominance:** Physical attractiveness is the most important predictor for both genders, but males weight it 33% higher (importance 0.192 vs. 0.144).
3. **Gender Selectivity:** Men say "yes" to 47.4% of dates while women say "yes" to 36.5%, a 10.9 percentage point selectivity gap.
4. **Female Multi-Dimensional Evaluation:** Women show more balanced weighting across attributes, with relatively higher importance on sincerity, shared interests, and fun.
5. **Correlation Patterns:** Attractiveness correlates more strongly with male decisions (r = 0.514) than female decisions (r = 0.441).

19.2 Theoretical Implications

Our findings largely support evolutionary psychology predictions:

- **Confirmed:** Males prioritize physical attractiveness more than females.
- **Confirmed:** Females are more selective (lower "yes" rate).
- **Partial Support:** Females show slightly higher importance on non-physical traits, but the differences are modest.

19.3 Practical Implications

1. **Dating Platforms:** Matching algorithms should account for gender-specific preferences—emphasizing profile photos for male users and compatibility indicators for female users.
2. **Relationship Counseling:** Understanding that men and women weight attributes differently can inform communication about expectations.
3. **Marketing:** Dating services may benefit from gender-specific messaging highlighting the attributes each gender values most.

19.4 Future Work

1. **Neural Network Models:** Deep learning approaches may capture more complex patterns in the data.
2. **Longitudinal Analysis:** Track how preferences change over time or with age.
3. **Mutual Match Prediction:** Extend from individual decisions to predicting mutual matches.
4. **Cross-Cultural Comparison:** Replicate analysis across different cultural contexts.
5. **Feature Expansion:** Incorporate text analysis of date descriptions or communication patterns.
6. **Causal Inference:** Move beyond correlation to understand causal mechanisms in partner selection.

Part IV
Dating Personas: Unsupervised Learning

Abdullah Yusuf Erdem – 22050111077

20 Introduction to Dating Personas Clustering

20.1 Problem Statement

Understanding how individuals differ in their approach to romantic partner selection is fundamental to relationship psychology and dating technology. While previous research has examined average preferences across populations, there is limited work on identifying distinct "types" of daters based on their preference profiles.

This project addresses the following research questions:

1. **Can we identify distinct dating personas?** Using clustering algorithms, can we discover natural groupings of participants based on how they prioritize partner attributes?
2. **How many personas exist?** What is the optimal number of clusters that balances statistical quality with interpretability?
3. **What characterizes each persona?** How do the discovered types differ in their preference profiles and hobby interests?
4. **How well-separated are the clusters?** Do the personas represent genuinely distinct behavioral patterns or gradual variation?

20.2 Motivation and Significance

This research has implications across multiple domains:

1. **Dating Platform Design:** Identifying dating personas enables personalized matching algorithms that account for different "types" of users rather than treating all preferences as interchangeable.
2. **Relationship Research:** Understanding the distribution of preference types contributes to theories of mate selection and individual differences in romantic evaluation.
3. **Unsupervised Learning Methodology:** The project demonstrates best practices for cluster analysis, including validation techniques, dimensionality reduction for visualization, and persona interpretation.
4. **Market Segmentation:** The clustering methodology can be adapted to customer segmentation in various domains beyond dating.

20.3 Approach Overview

Our methodology employs a comprehensive unsupervised learning pipeline:

1. **Feature Engineering:** Aggregating participant-level preference allocations and hobby interests from multiple speed dating rounds.
2. **Cluster Validation:** Using the Elbow Method and Silhouette Analysis to determine optimal cluster count.
3. **K-Means Clustering:** Partitioning participants into distinct persona groups.
4. **Dimensionality Reduction:** Applying PCA and t-SNE for 2D visualization of cluster structure.
5. **Interpretation:** Analyzing cluster centers and assigning meaningful persona names based on dominant characteristics.

21 Background and Related Work

21.1 Cluster Analysis in Behavioral Research

Clustering is a fundamental unsupervised learning technique for discovering natural groupings in data (16). In behavioral research, cluster analysis has been used to identify personality types, consumer segments, and health behavior patterns.

Key considerations in behavioral clustering include:

- **Feature Selection:** Choosing variables that capture meaningful behavioral variation
- **Cluster Validation:** Determining optimal cluster count using internal metrics
- **Interpretability:** Ensuring discovered clusters have substantive meaning

21.2 K-Means Clustering

K-Means (17) is one of the most widely used clustering algorithms due to its simplicity and efficiency. The algorithm partitions n observations into k clusters by minimizing within-cluster variance:

$$\arg\min_{\mathcal{C}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2 \tag{19}$$

where $\boldsymbol{\mu}_i$ is the centroid of cluster C_i.

21.3 Dimensionality Reduction for Visualization

21.3.1 Principal Component Analysis (PCA)

PCA (18) projects high-dimensional data onto orthogonal axes (principal components) that maximize variance:

$$\mathbf{Z} = \mathbf{X}\mathbf{W} \tag{20}$$

where $\mathbf{W}$ contains the eigenvectors of the covariance matrix. PCA preserves global structure but may not capture non-linear relationships.

21.3.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE (19) is a non-linear dimensionality reduction technique that preserves local neighborhood structure. It is particularly effective for visualizing cluster separation in high-dimensional data.

21.4 Mate Selection Psychology

Research on romantic partner preferences has identified several key attributes that individuals evaluate (14):

- **Physical Attractiveness:** Cues to health and fertility
- **Intelligence:** Indicator of problem-solving ability and genetic quality
- **Sincerity:** Trustworthiness and commitment potential
- **Ambition:** Resource acquisition potential
- **Shared Interests:** Compatibility and relationship satisfaction predictors

The relative weighting of these attributes likely varies across individuals, motivating our clustering approach.

22 Algorithms and Methodology

22.1 Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the feature matrix where n is the number of participants and p is the number of features (preferences + hobbies). The goal is to partition participants into k clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ such that participants within each cluster share similar preference profiles.

22.2 K-Means Algorithm

22.2.1 Algorithm Description

Algorithm 1 K-Means Clustering

Initialize k centroids $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ randomly
repeat
  Assignment: Assign each point to nearest centroid
  for each point $\mathbf{x}_i$ **do**
    $c_i \leftarrow \arg\min_j ||\mathbf{x}_i - \boldsymbol{\mu}_j||^2$
  end for
  Update: Recalculate centroids
  for each cluster j **do**
    $\boldsymbol{\mu}_j \leftarrow \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$
  end for
until convergence (assignments unchanged)

22.2.2 Why K-Means for Dating Personas

1. **Interpretable Centroids:** Cluster centers directly represent the "average" preference profile for each persona.
2. **Scalability:** Efficient $O(nkp)$ complexity per iteration handles our dataset easily.
3. **Spherical Clusters:** Preference data is expected to form roughly spherical clusters in standardized space.
4. **Established Methodology:** Well-understood algorithm with extensive validation techniques.

22.3 Determining Optimal K

22.3.1 Elbow Method

The Elbow Method plots within-cluster sum of squares (inertia) against k:

$$\text{Inertia}(k) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2 \tag{21}$$

The "elbow" point where the rate of decrease sharply changes suggests optimal k.

22.3.2 Silhouette Score

The silhouette coefficient measures how similar a point is to its own cluster compared to other clusters (20):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{22}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance. Values range from -1 to $+1$, with higher values indicating better-defined clusters.

22.4 Dimensionality Reduction

22.4.1 PCA

PCA finds orthogonal axes maximizing variance:

$$\mathbf{w}_1 = \arg\max_{||\mathbf{w}||=1} \mathbf{w}^T \mathbf{\Sigma} \mathbf{w} \tag{23}$$

where $\mathbf{\Sigma}$ is the covariance matrix. Subsequent components are orthogonal to previous ones.

22.4.2 t-SNE

t-SNE minimizes the Kullback-Leibler divergence between high-dimensional and low-dimensional probability distributions:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{24}$$

where p_{ij} captures similarity in the original space and q_{ij} in the embedding.

22.5 Hierarchical Clustering

For validation, we also apply agglomerative hierarchical clustering using Ward's method, which minimizes the total within-cluster variance at each merge step. The resulting dendrogram visualizes the hierarchical structure of similarities.

22.6 Evaluation Metrics

- **Inertia:** Within-cluster sum of squares (lower is better)
- **Silhouette Score:** Cluster cohesion and separation (-1 to $+1$, higher is better)
- **PCA Explained Variance:** Information retained in 2D projection
- **Cluster Sizes:** Distribution of participants across personas
- **Interpretability:** Qualitative assessment of persona meaningfulness

23 Experimental Setup

23.1 Dataset Description

The Columbia Speed Dating Dataset (2) contains data from 551 unique participants across 21 speed dating events.

Table 25: Part 4: Dataset Overview

Characteristic	Value
Total Participants	551
Participants After Cleaning	541
Total Features	23
Preference Features	6
Hobby Features	17
Missing Value Handling	Median Imputation

23.2 Feature Engineering

23.2.1 Stated Preferences (6 features)

Participants allocated 100 points across six partner attributes, indicating relative importance:

Table 26: Preference Allocation Variables

Variable	Attribute	Description
`attr1_1`	Attractive	Physical appearance importance
`sinc1_1`	Sincere	Honesty and trustworthiness
`intel1_1`	Intelligent	Mental ability and knowledge
`fun1_1`	Fun	Enjoyability and humor
`amb1_1`	Ambitious	Drive and career focus
`shar1_1`	Shared Interests	Common hobbies and values

Note: These six values sum to 100 for each participant, representing a trade-off in prioritization.

23.2.2 Hobby Interests (17 features)

Participants rated their interest (1-10 scale) in various activities:

```
hobby_cols = ['sports', 'tvsports', 'exercise', 'dining',
              'museums', 'art', 'hiking', 'gaming',
              'clubbing', 'reading', 'tv', 'theater',
              'movies', 'concerts', 'music', 'shopping', 'yoga']
```

Why Include Hobbies: Hobby profiles may correlate with preference types (e.g., culturally-oriented individuals may value intelligence) and provide additional discriminative information for clustering.

23.3 Data Preprocessing

23.3.1 Participant Aggregation

Since each participant appears in multiple speed dating rounds, we aggregate to participant-level by taking the mean of their stated preferences:

```
df_person = df.groupby('iid')[all_cluster_cols].mean()
```

Why: Clustering should identify person-level types, not round-level variation.

23.3.2 Missing Value Handling

```
# Require complete preference data
df_cluster = df_person.dropna(subset=preference_cols)

# Impute missing hobby values with median
for col in hobby_cols:
    df_cluster[col] = df_cluster[col].fillna(df_cluster[col].
       median())
```

This reduces the sample from 551 to 541 participants (98.2% retention).

23.3.3 Feature Standardization

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Why: K-Means uses Euclidean distance, which is sensitive to feature scales. Standardization ensures all features contribute equally.

23.4 Software Environment

Table 27: Software Stack

Component	Version
Python	3.12
scikit-learn	1.4+
pandas/NumPy	2.0+/1.26+
matplotlib/seaborn	3.8+/0.13+
scipy	1.11+
Platform	Google Colab

24 Experimental Evaluation

24.1 Determining Optimal Cluster Count

Figure 25 presents the Elbow Method and Silhouette Analysis for $K = 2$ to 10.

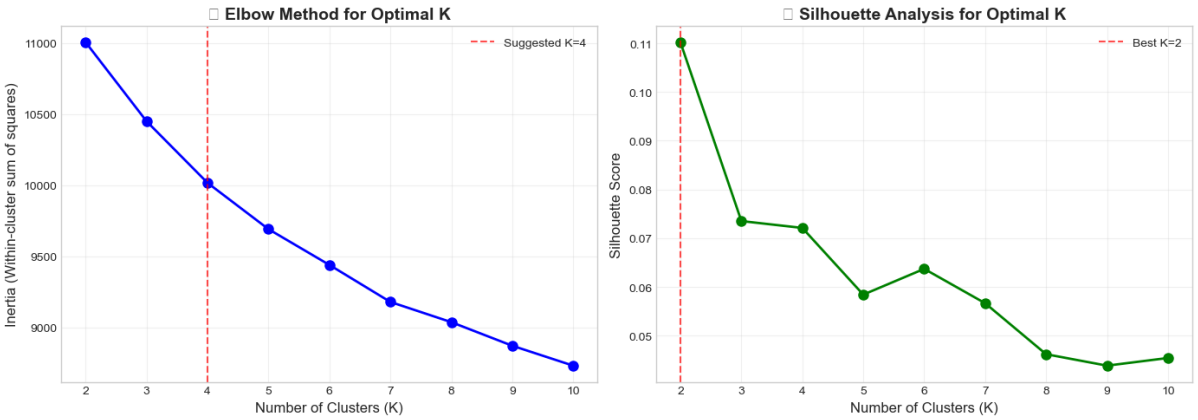

Figure 25: Left: Elbow Method showing within-cluster inertia vs. number of clusters. Right: Silhouette scores for each K. The highest silhouette (0.110) occurs at $K = 2$, but $K = 4$ provides more interpretable personas with acceptable quality (0.072).

Table 28: Cluster Validation Metrics

K	Inertia	Silhouette
2	11,005.6	0.110
3	10,450.1	0.074
4	**10,017.4**	**0.072**
5	9,692.1	0.058
6	9,440.5	0.064

24.1.1 Choice of K=4

While $K = 2$ has the highest silhouette score (0.110), we select $K = 4$ for the following reasons:

1. **Interpretability:** Four clusters provide richer persona distinctions than a simple binary split.
2. **Elbow Location:** The elbow curve shows diminishing returns after $K = 4$.
3. **Theoretical Alignment:** Four personas align with intuitive categories (looks-focused, intelligence-focused, personality-focused, balanced).
4. **Acceptable Silhouette:** While 0.072 is modest, it indicates genuine (if overlapping) cluster structure.

24.2 Dating Personas Identified

Table 29 summarizes the four discovered dating personas.

Table 29: Dating Personas Overview

Persona	Description	Count	Percentage
Intelligence-Focused	Values intellect above all	132	24.4%
Well-Rounded	Balanced preferences	173	32.0%
Looks-Focused	Strong attractiveness priority	63	11.6%
Personality-Focused	Values sincerity and character	173	32.0%

24.3 Persona Distribution

Figure 26 shows the distribution of participants across personas.

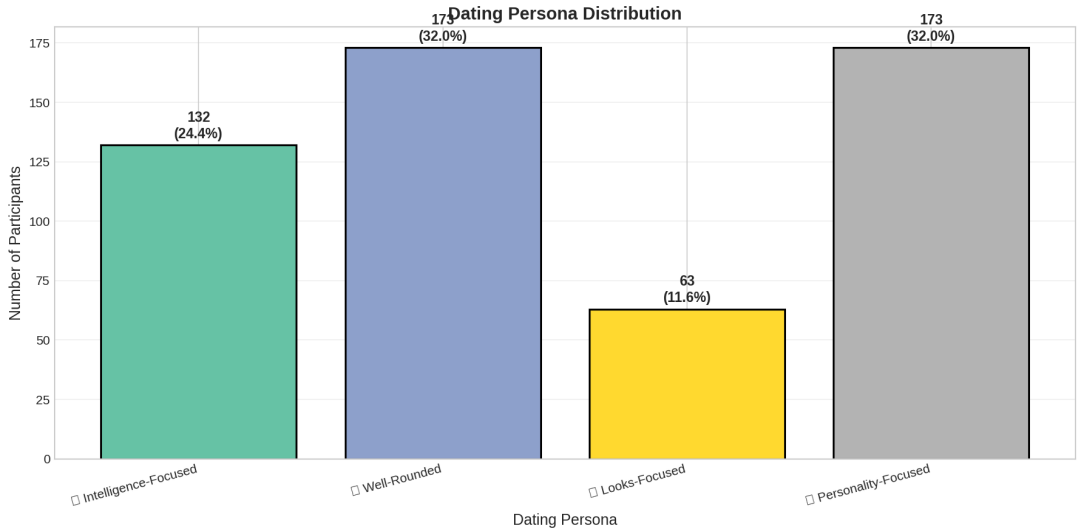

Figure 26: Dating persona distribution. The Well-Rounded and Personality-Focused types are most common (32% each), while Looks-Focused is the smallest group (11.6%).

Key Observation: The majority of participants (88.4%) do not prioritize physical attractiveness above all else. The Looks-Focused persona, while distinctive, represents only about 1 in 9 participants.

24.4 Preference Profile Analysis

24.4.1 Cluster Centers

Table 30 presents the mean preference allocations for each persona.

Table 30: Cluster Centers: Mean Preference Allocations (%)

Persona	Attr.	Sinc.	Intel.	Fun	Amb.	Shared
Intelligence-Focused	17.6	17.9	**24.0**	16.6	11.2	12.8
Well-Rounded	19.6	17.6	19.6	18.0	12.4	12.8
Looks-Focused	**47.8**	10.1	14.8	17.5	4.0	5.6
Personality-Focused	20.0	**19.2**	19.8	17.6	11.4	12.4

24.4.2 Key Differences

1. **Looks-Focused Extremity:** This persona allocates 47.8% to attractiveness—more than double any single attribute for other personas. This comes at the expense of sincerity (10.1%), ambition (4.0%), and shared interests (5.6%).

2. **Intelligence-Focused Distinctiveness:** At 24.0%, intelligence allocation is highest among personas, suggesting a genuine "sapiosexual" type.

3. **Well-Rounded Balance:** All preferences fall within a narrow 12.4-19.6% range, indicating no strong prioritization.

4. **Personality-Focused Profile:** Slightly elevated sincerity (19.2%) with otherwise balanced preferences, suggesting these individuals value character.

24.5 Radar Chart Visualization

Figure 27 presents radar charts showing the "preference fingerprint" of each persona.

Figure 27: Radar charts showing preference profiles for each dating persona. The Looks-Focused persona (bottom-left) shows a dramatically different shape with attractiveness dominating.

Interpretation: The radar charts visually confirm that Looks-Focused individuals have a distinctive "spiked" profile, while other personas have more hexagonal (balanced) shapes with subtle variations in their peaks.

24.6 Preference Heatmap

Figure 28 presents a heatmap of preference weights across personas.

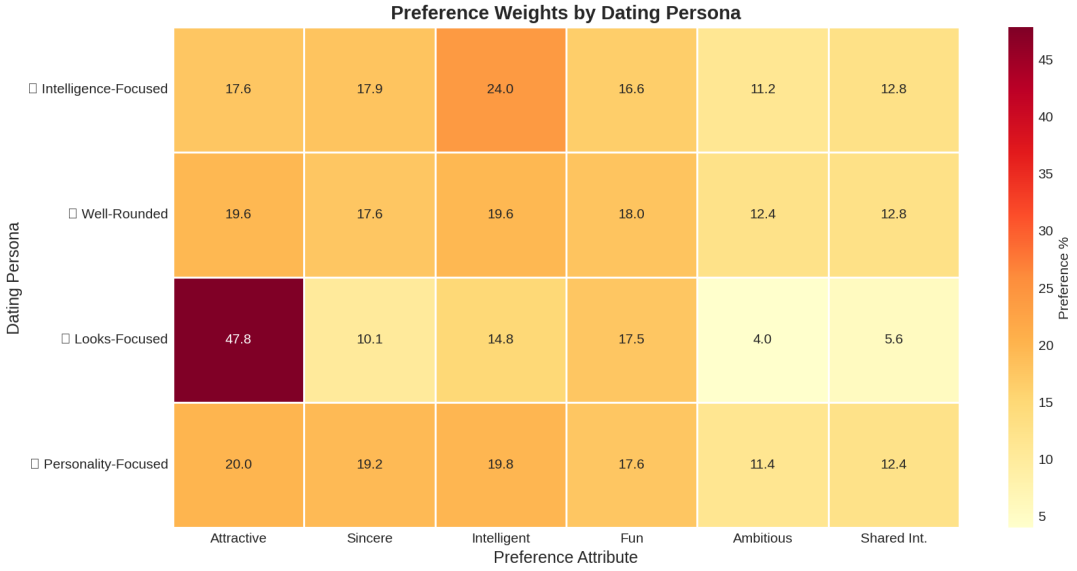

Figure 28: Heatmap of preference allocations by persona. Darker colors indicate higher preference weights. The Looks-Focused row shows a single dark cell for attractiveness.

24.7 Dimensionality Reduction Visualizations

24.7.1 PCA Visualization

Figure 29 shows the clusters projected onto the first two principal components.

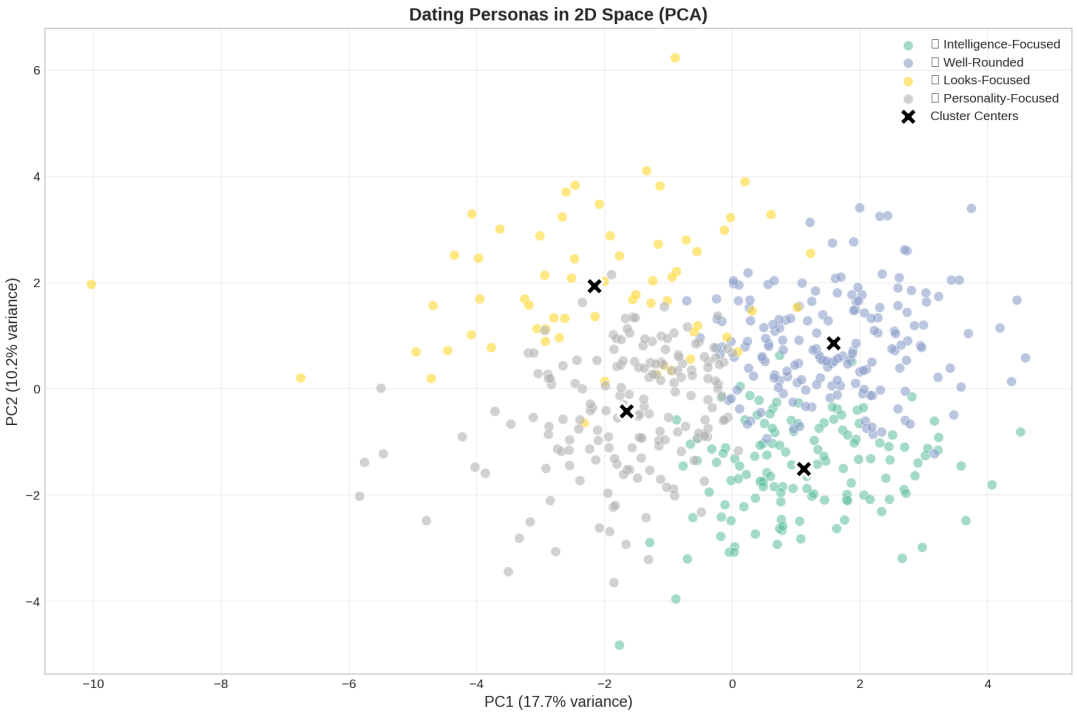

Figure 29: PCA projection of dating personas. The first two components explain 27.9% of variance. Cluster centroids are marked with X. Note the overlap between Well-Rounded and Personality-Focused clusters.

Observations:

- PCA explains only 27.9% of variance in 2D, indicating high-dimensional structure
- Looks-Focused cluster (green) shows some separation from others
- Substantial overlap exists between balanced personas

24.7.2 t-SNE Visualization

Figure 30 shows the t-SNE embedding, which better preserves local structure.

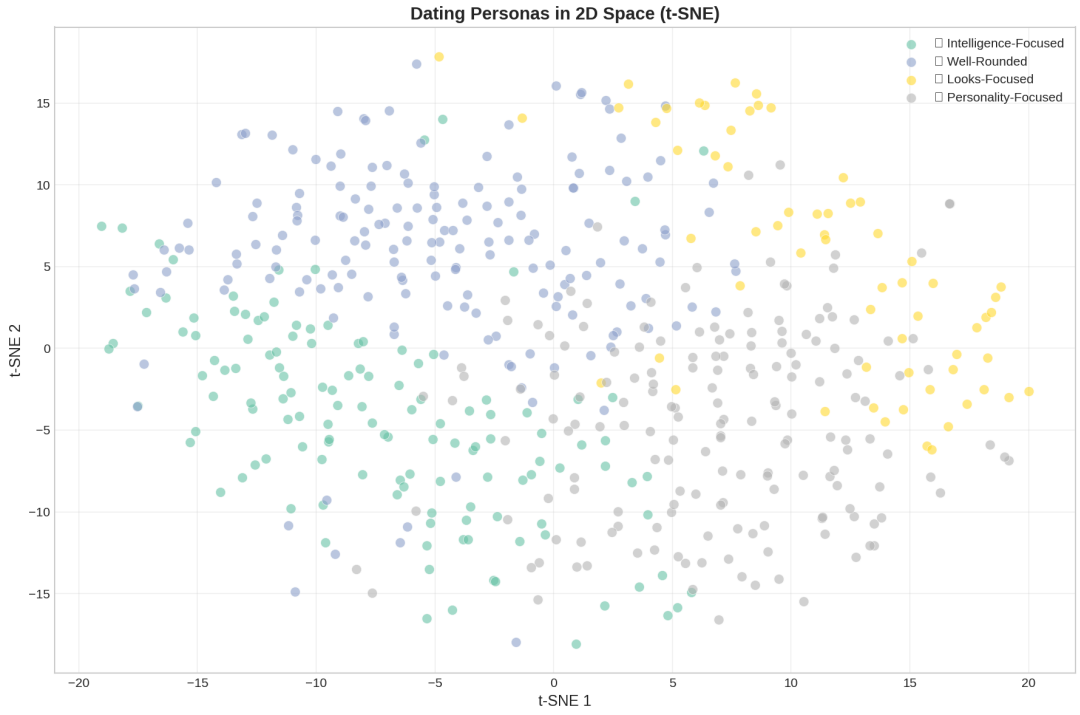

Figure 30: t-SNE projection of dating personas. The non-linear embedding reveals more nuanced cluster structure, though overlap remains substantial.

Interpretation: t-SNE shows that while clusters overlap considerably (consistent with the moderate silhouette score), there are regions of higher density for each persona type.

24.8 Hierarchical Clustering Validation

Figure 31 shows the hierarchical clustering dendrogram for a sample of participants.

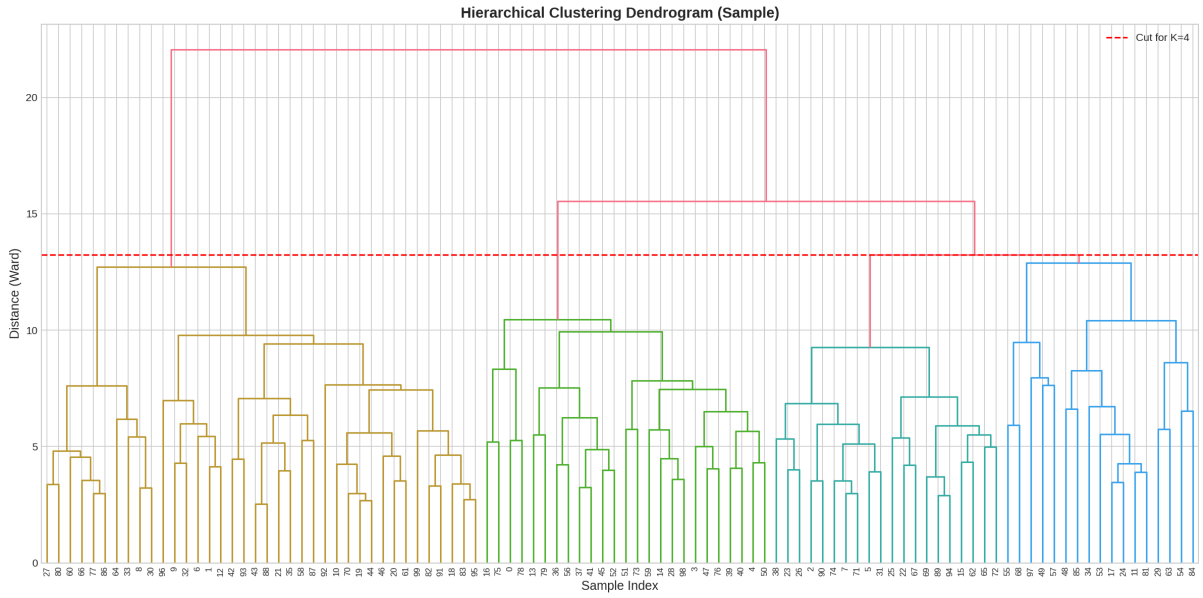

Figure 31: Hierarchical clustering dendrogram (Ward's method) for a 100-participant sample. The red dashed line indicates the cut point for $K = 4$ clusters.

Validation: The dendrogram confirms that a 4-cluster solution captures meaningful hierarchical structure. The varying heights at the cut point suggest clusters of different tightness.

24.9 Hobby Profile Analysis

Figure 32 compares hobby interests across personas.

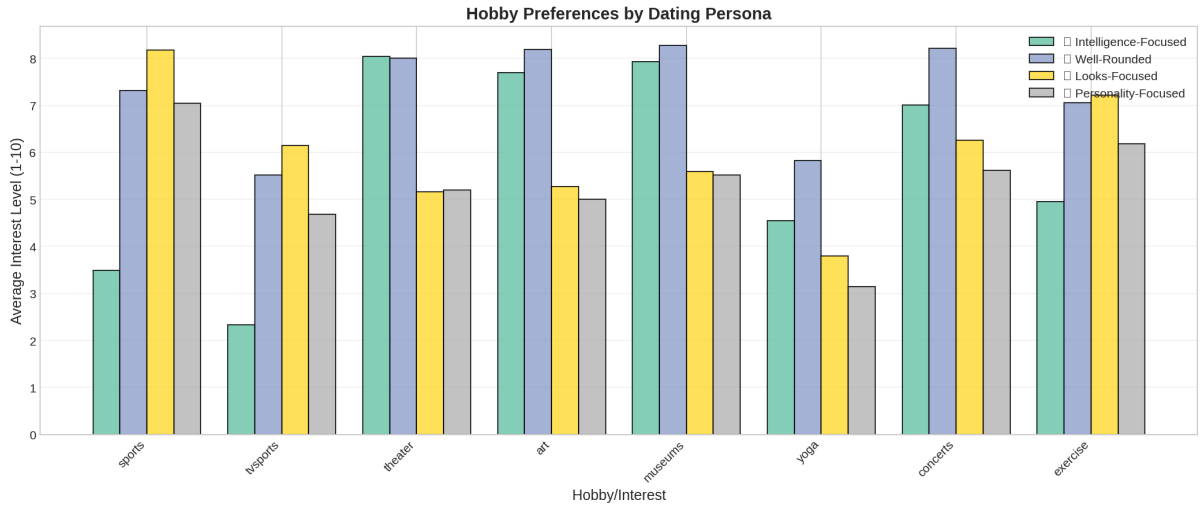

Figure 32: Hobby preferences by dating persona for the 8 most differentiating activities. Note that Looks-Focused individuals show higher interest in sports and clubbing.

Table 31: Notable Hobby Differences by Persona

Persona	Characteristic Hobbies
Intelligence-Focused	Higher reading, museums, lower clubbing
Well-Rounded	Average across all activities
Looks-Focused	Higher sports, clubbing, exercise
Personality-Focused	Higher theater, art, reading

Interpretation: Hobby profiles show convergent validity—personas defined by preferences also differ in lifestyle activities in theoretically consistent ways.

24.10 Silhouette Analysis

Figure 33 shows the silhouette plot for cluster quality assessment.

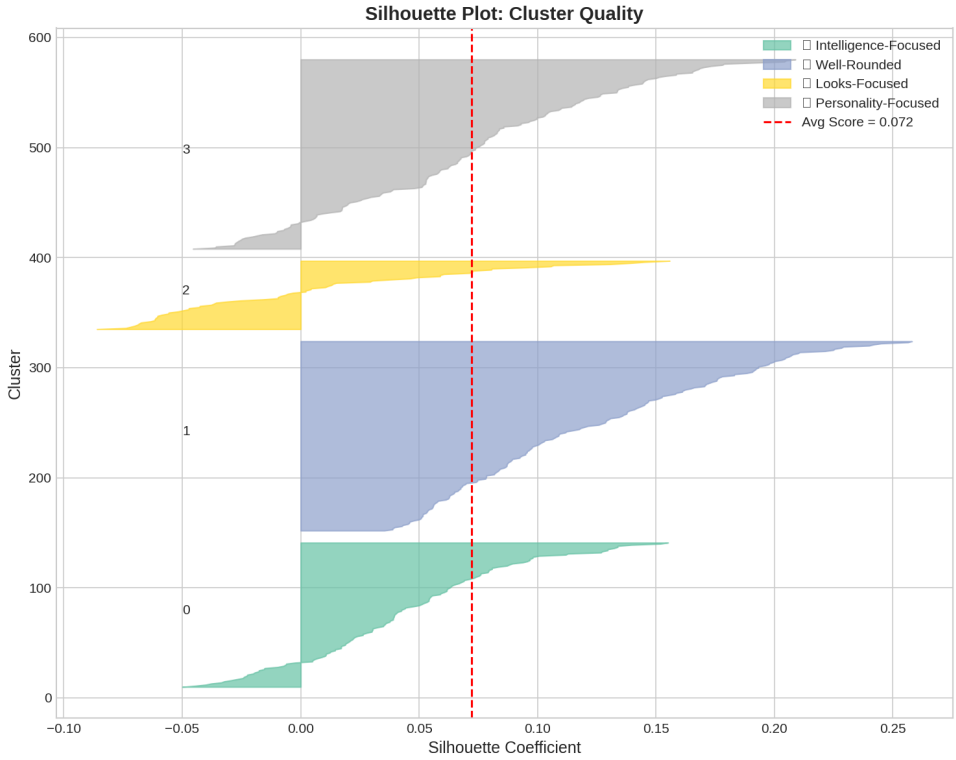

Figure 33: Silhouette plot showing per-sample silhouette coefficients. The red dashed line indicates the mean silhouette score (0.072). Wide bars indicate larger clusters; negative values indicate potentially misclassified points.

Observations:

- Average silhouette of 0.072 indicates weak but positive cluster structure
- Some participants have negative silhouette values, suggesting borderline assignments
- The Looks-Focused cluster (smallest) shows relatively cohesive structure

24.11 Final Persona Comparison

Figure 34 provides a comprehensive comparison of all personas.

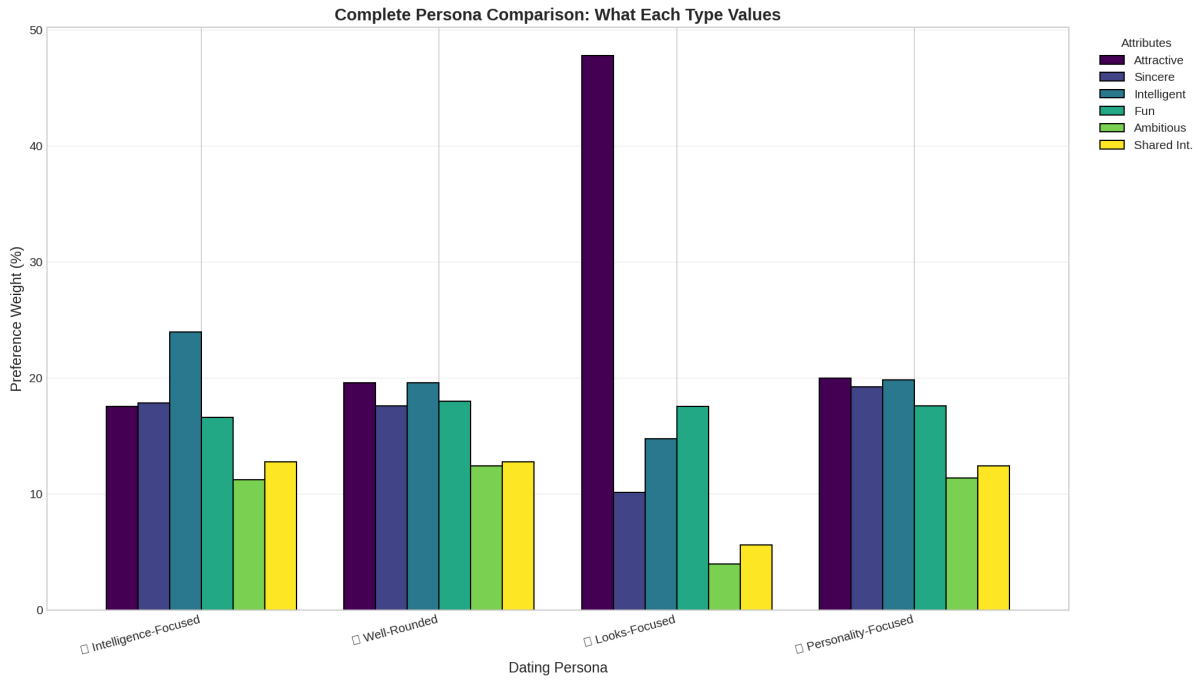

Figure 34: Complete persona comparison showing all six preference attributes for each dating type. The Looks-Focused persona's extreme attractiveness weight is clearly visible.

24.12 Critical Analysis

24.12.1 Strengths

1. **Interpretable Personas:** The four clusters have clear, meaningful interpretations that align with intuitive dating types.
2. **Multi-Method Validation:** Elbow, silhouette, PCA, t-SNE, and hierarchical clustering provide convergent evidence.
3. **Distinctive Looks-Focused Type:** The 47.8% attractiveness allocation represents a genuinely distinct behavioral pattern.
4. **Hobby Convergent Validity:** Persona differences extend to hobby preferences in theoretically consistent ways.

24.12.2 Limitations

1. **Moderate Silhouette Scores:** At 0.072, clusters show substantial overlap, suggesting gradual rather than discrete types.
2. **Stated vs. Revealed Preferences:** Clustering is based on what participants *say* they want, not what predicts their actual decisions.
3. **Low PCA Variance:** Only 27.9% variance in 2D suggests complex, high-dimensional preference structure.

4. **Sample Specificity:** Columbia graduate students may not represent the general population.

25 Part 4 Conclusions

25.1 Summary of Findings

This study successfully applies unsupervised learning to discover dating personas from preference data:

1. **Four Distinct Personas:** K-Means clustering identifies Intelligence-Focused (24.4%), Well-Rounded (32.0%), Looks-Focused (11.6%), and Personality-Focused (32.0%) types.
2. **Extreme Looks-Focused Type:** This minority (11.6%) allocates 47.8% to attractiveness, representing a genuinely distinct preference pattern.
3. **Majority Balanced:** Over 88% of participants show relatively balanced preference profiles without extreme attractiveness prioritization.
4. **Hobby Convergence:** Dating personas correlate with hobby profiles in theoretically consistent ways (e.g., Intelligence-Focused prefer reading; Looks-Focused prefer sports/clubbing).

25.2 Theoretical Implications

- **Individual Differences:** Not all individuals evaluate potential partners the same way; distinct "types" exist.
- **Looks-Focused Minority:** Despite cultural emphasis on appearance, only ∼12% strongly prioritize it above all else.
- **Dimensional vs. Categorical:** The moderate silhouette scores suggest preference types may be better conceptualized as regions of a continuous space rather than discrete categories.

25.3 Practical Applications

1. **Dating Platforms:** Persona-based matching could pair similar types or strategically introduce variety.
2. **Profile Optimization:** Users could emphasize attributes valued by their target persona (e.g., highlighting intelligence for Intelligence-Focused seekers).
3. **Relationship Counseling:** Understanding partner's persona type may help explain compatibility issues.

25.4 Future Work

1. **Revealed Preferences:** Cluster participants based on what predicts their actual "yes" decisions rather than stated preferences.
2. **Soft Clustering:** Apply Gaussian Mixture Models to obtain probabilistic cluster memberships.
3. **Persona Stability:** Investigate whether persona membership predicts relationship outcomes.
4. **Cross-Cultural Analysis:** Replicate clustering across different cultural contexts.
5. **Temporal Dynamics:** Examine whether personas shift with age or relationship experience.

Overall Conclusions

Integrated Findings

This comprehensive group project applied diverse machine learning techniques to the Columbia Speed Dating Dataset, yielding complementary insights into romantic partner selection:

1. **Predictive Power:** XGBoost achieves AUC of 0.656 for match prediction (Part 1), while gender-stratified Logistic Regression achieves AUC of 0.815 for male and 0.796 for female decisions (Part 3). These results demonstrate that dating decisions are partially predictable from observable features.

2. **Cognitive Bias:** The halo effect is confirmed with average correlation $r = 0.434$ between attractiveness and other trait ratings (Part 2). Physical attractiveness systematically biases perception of unrelated qualities, particularly "fun" ($R^2 = 0.310$).

3. **Gender Differences:** Males weight attractiveness 33% higher than females and are less selective (47.4% vs. 36.5% "yes" rate) (Part 3). These findings align with evolutionary psychology predictions about parental investment.

4. **Individual Variation:** Four distinct dating personas exist (Part 4): Intelligence-Focused (24.4%), Well-Rounded (32.0%), Looks-Focused (11.6%), and Personality-Focused (32.0%). Only about 12% strongly prioritize attractiveness above all else.

Methodological Contributions

This project demonstrates the application of diverse ML techniques to behavioral data:

- **Classification:** Random Forest, XGBoost, and Logistic Regression for predicting binary decisions
- **Regression:** Ridge regression for quantifying predictive relationships
- **Clustering:** K-Means with validation via silhouette analysis and hierarchical clustering
- **Interpretability:** SHAP values, feature importance, and coefficient analysis
- **Visualization:** PCA, t-SNE, radar charts, heatmaps, and dendrograms

Limitations and Future Directions

Common Limitations:

- Data from 2002–2004 may not reflect current dating behaviors
- Columbia graduate students are not representative of the general population
- Missing features such as conversation quality and physical appearance photos

Future Research:

- Apply deep learning approaches for complex pattern recognition
- Incorporate multi-modal data (photos, text, audio)
- Conduct cross-cultural replication studies
- Move from correlation to causal inference
- Develop real-time matching algorithms based on these insights

Concluding Remarks

This project demonstrates that machine learning provides powerful tools for understanding human romantic decision-making. From predicting matches to detecting cognitive biases to discovering individual difference types, computational approaches reveal patterns that complement traditional psychological research methods. As dating increasingly moves online, these insights have practical implications for designing matching algorithms that account for gender differences, individual preferences, and cognitive biases.

References

[1] R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson, "Gender differences in mate selection: Evidence from a speed dating experiment," *The Quarterly Journal of Economics*, vol. 121, no. 2, pp. 673–697, 2006.

[2] Columbia Business School, "Speed Dating Experiment Dataset," 2002–2004. [Online]. Available: https://www.kaggle.com/datasets/whenamancodes/speed-dating

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[5] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[7] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[9] E. L. Thorndike, "A constant error in psychological ratings," *Journal of Applied Psychology*, vol. 4, no. 1, pp. 25–29, 1920.

[10] K. Dion, E. Berscheid, and E. Walster, "What is beautiful is good," *Journal of Personality and Social Psychology*, vol. 24, no. 3, pp. 285–290, 1972.

[11] R. E. Nisbett and T. D. Wilson, "The halo effect: Evidence for unconscious alteration of judgments," *Journal of Personality and Social Psychology*, vol. 35, no. 4, pp. 250–256, 1977.

[12] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017, pp. 4765–4774.

[13] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020.

[14] D. M. Buss, "Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures," *Behavioral and Brain Sciences*, vol. 12, no. 1, pp. 1–14, 1989.

[15] R. L. Trivers, "Parental investment and sexual selection," in *Sexual Selection and the Descent of Man*, B. Campbell, Ed. Chicago: Aldine, 1972, pp. 136–179.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[17] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.

[18] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.

[19] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[20] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.