Group3_EDA_ProjectPlan

AUTHOR
Daniel Rishea, Maria Granados, Mosaab Saleem

Romantic Interest in Speed Dating

Introduction

This project investigates the attributes that contribute to *romantic interest* in a speed dating context. The dataset, obtained from OpenML (Speed Dating Dataset; OpenML ID 40536), contains outcomes of speed dating events conducted between 2002 and 2004, including lifestyle and demographic information about both participants.

Previous studies have used this dataset to predict "matches" between participants (Fisman et al., 2006). However, we noted that the experimental design favoured the subject of the date, which provided an unequal number of features compared to their partner. Consequently, we reframed our problem: instead of predicting matches, we aim to predict the **partner's decision** (binary variable: *decision* = yes/no) regarding whether they would like to meet the subject again.

Specifically, we seek to:

1. Examine the relationship between **self-reported subject attributes** (how individuals describe themselves) and **partner perceptions** (how those attributes are rated by the partner).
2. Compare how these factors differ across **gender**.
3. Identify which features most strongly predict *romantic interest* in this context.

Exploratory Data Analysis

We first completed data cleaning and exploratory data analysis (EDA). This step was essential to ensure we could extract insights from the dataset, to guide the choice of appropriate models and evaluation metrics in line with our problem statement.

Dataset Description

The original data contained 121 features and 8,378 observations. However, 87.4% of rows contained at least one missing value. The original dataset presented several challenges:

- High number of missing values in some features.
- High dimensionality with many overlapping features, resulting in high correlations between many features and target variable.
- Unclear feature nomenclature. This was a considerable challenge, which required an iterative approach to cross-reference variables with documentation and related literature. Although time-intensive, this was essential to select meaningful features that aligned with our research question.

Data Cleaning Methodology

To prepare the data for Exploratory Data Analysis (EDA), we completed the following steps:

- Cross-referenced variables with documentation to keep only relevant variables:
  - Removed feature ranges when raw numeric values were already available.
  - To avoid data leakage, features directly correlated with the target variable were removed (e.g., "match," "interest correlation").
  - Dropped irrelevant columns (e.g., "expected," "happy," "prob_liked").
- After filtering features, we removed rows with missing values in the remaining dataset. We took this approach to ensure we were retaining as many relevant observations as possible, as the original dataset had 80% missing observations.
  - Overall missing values for the relevant features were 57.7%.
  - When calculated by column, this was a considerable improvement, with the top three fields with missing values showing less than 15% nulls present.

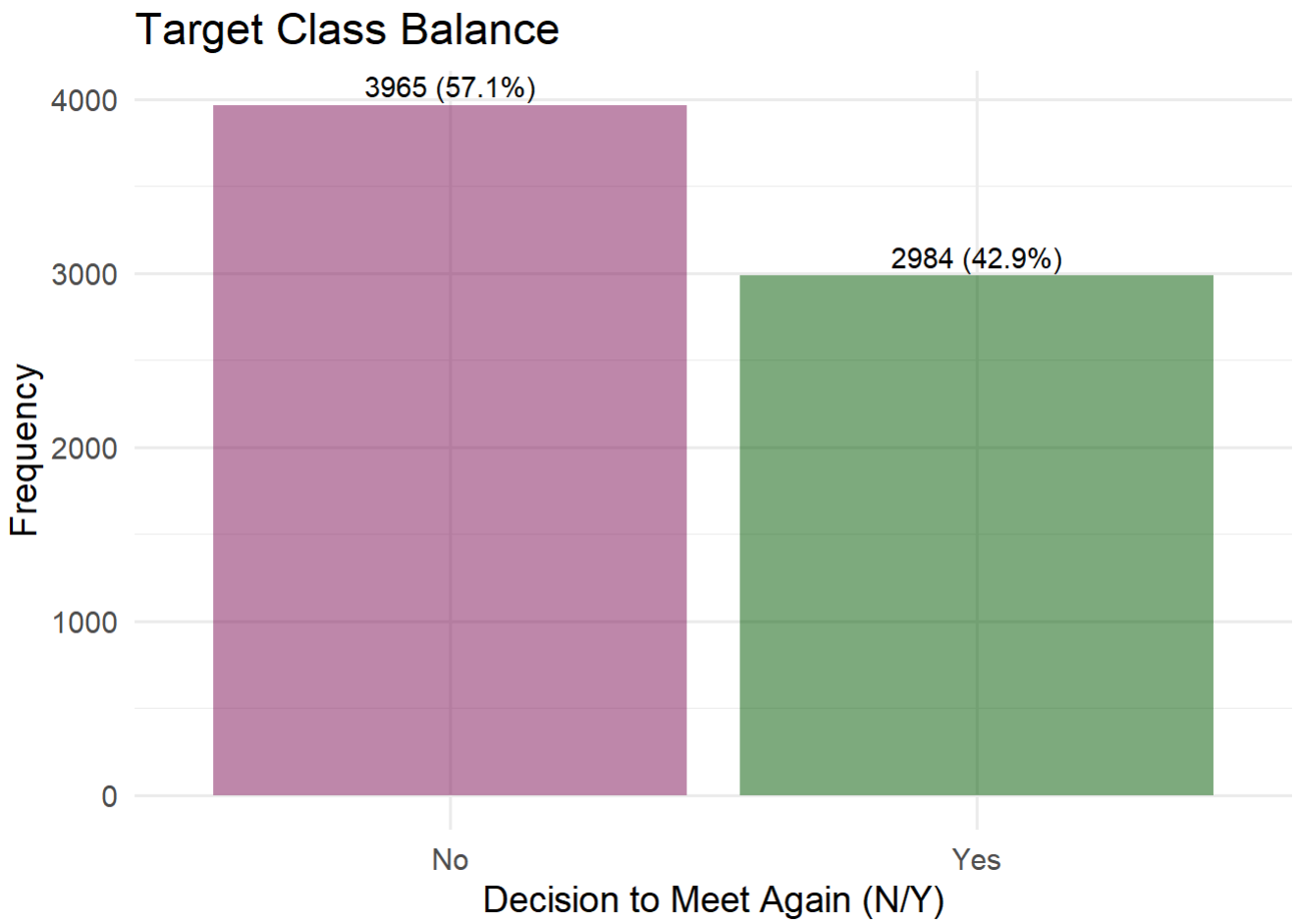
Feature	NullCount	NullPercent
shared_interests_o	1076	12.84
ambitous_o	722	8.62
funny_o	360	4.30

- Renamed target variable to *decision* for clarity.
- Out-of-range values (expected 1–10 scale) were handled using three strategies:
  1. Retained values as-is,
  2. Removed out-of-range values,
  3. Rescaled using min–max transformation.

For the EDA, we retained values as-is to assess their distribution and impact before finalising an approach. Our final cleaned dataset included **33 features and 6,949 observations** (see Appendix for details on the final dataset).

Target Variable Balance

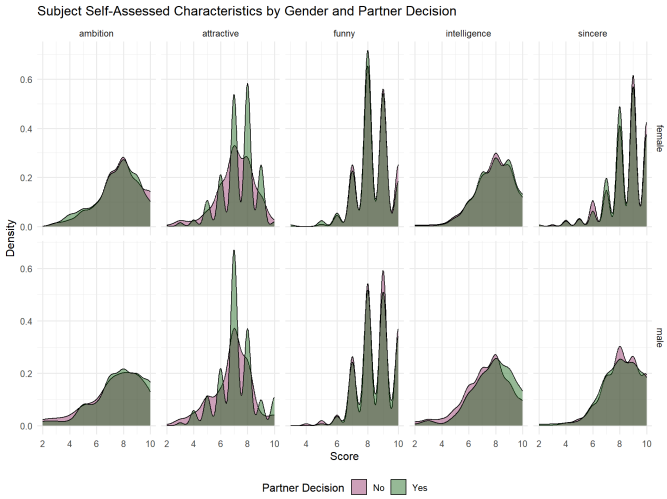
The dataset is relatively balanced as shown below. As class imbalance is not severe, the risk of class imbalance bias is low which supports the use of accuracy as a baseline evaluation metric, but nuanced metrics (F1, ROC-AUC) will still be important.

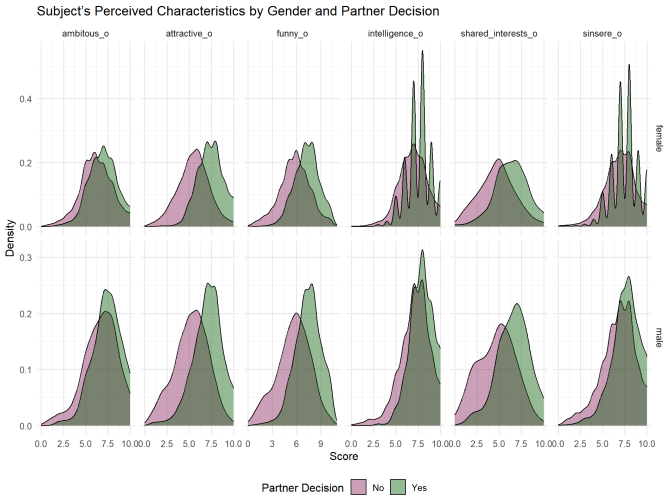


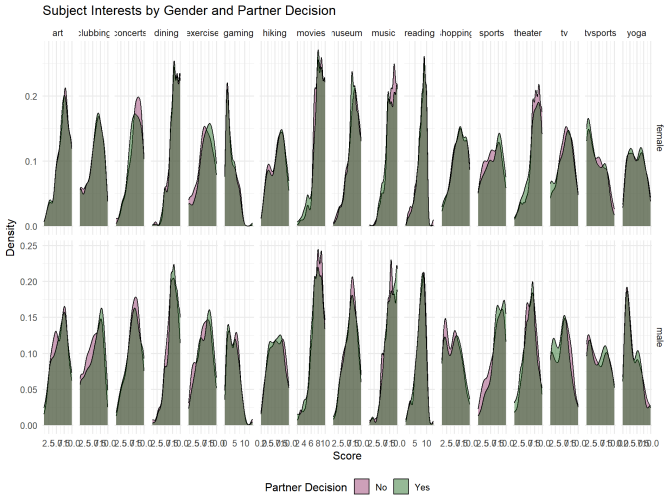
Feature Distributions and Trends

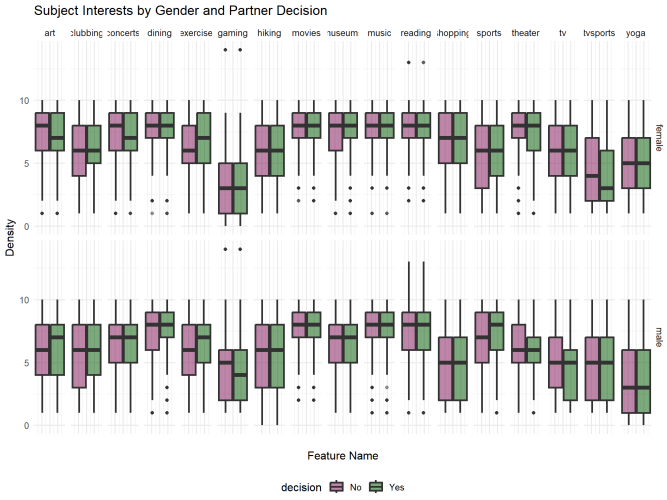
Key insights analysis show:

- **Gender and characteristics interaction:**
    - **Subject self-assessed characteristics:** There is considerable overlap across most traits, with the exception of female partners seemingly favouring male subjects that self-report highly for attractiveness and intelligence, whilst male partners are seemingly less picky. It's interesting to point out that self-ratings by the subject do not seem to have a high effect on the final decision for male partner / female subject pairs, but yes for the female partners..
    - **Subject's perceived characteristics**: The subject's perceived characteristics by the partner show stronger differentiation, with perceived attractiveness, funny, and shared interest providing a clear indication of a favourable decision. There are some gender differences, with intelligence being a weaker indicator for male partner / female subject pairs, whilst female partners seem to highly regard this trait in comparison to male partners.
    - **Subject's interests**: The subject's interests also have considerable overlaps, with shopping, hiking, gaming and exercise showing the strongest differentiation. Interestingly, female partners seem to have less favorable views to theatre and tv as interests for male subjects, whilst male partners seemingly having unfavourable views to females with tv sports as an interest.

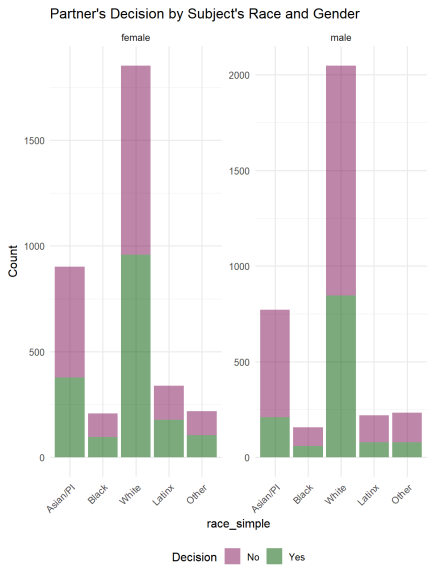


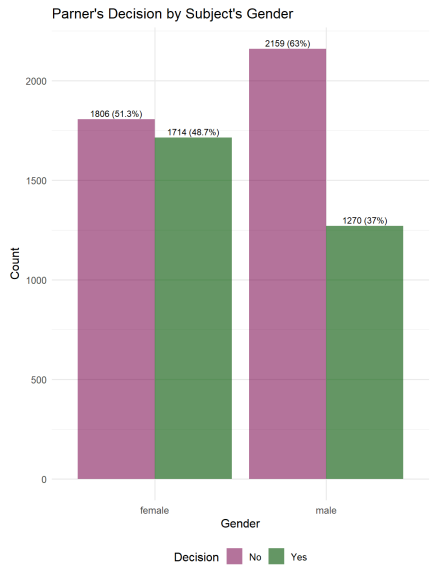






- **Gender and race interaction**:
  - Same-race pairings were more likely to result in "yes" decisions. Male subjects tended to receive more "no" responses overall.
  - Male partners showed favourable bias towards female subjects of White or Latina backgrounds, while other groups faced lower acceptance.
  - Female partners, appeared more race-conscious when evaluating male subjects, with Asian-American male subjects received a higher proportion of unfavourable responses.

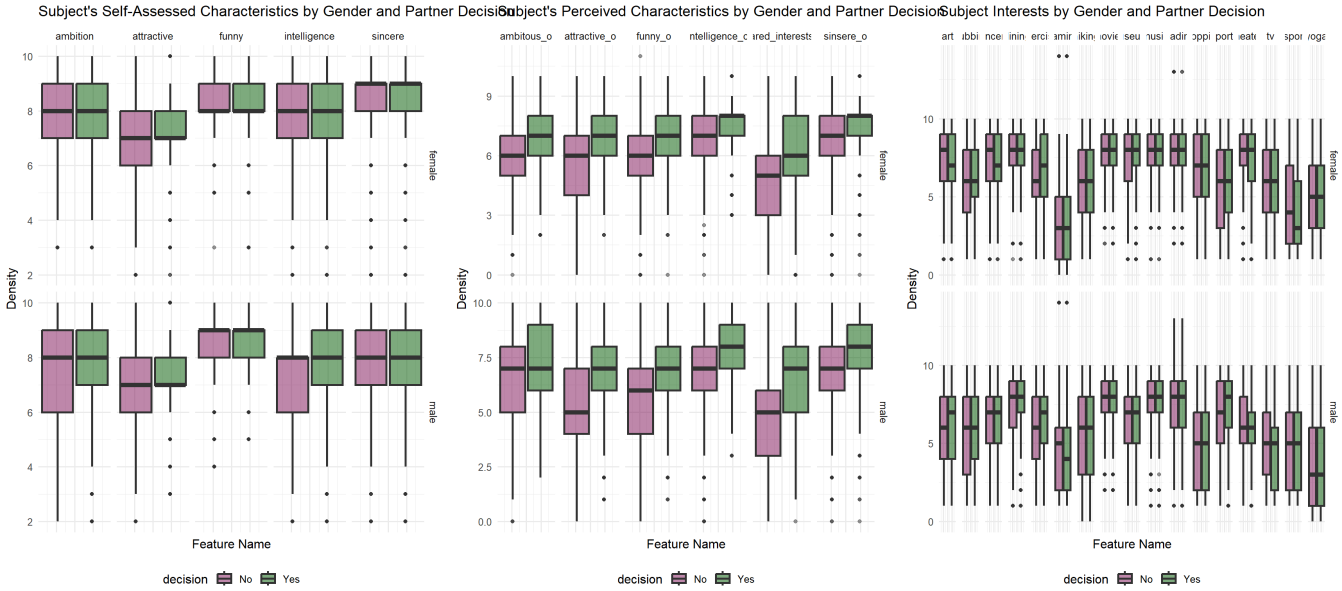




Outliers

Boxplots reveal several potential outliers beyond the interquartile range, especially in self reported ratings (1-10 scales with values outside the range).

- Models sensitive to outliers (e.g., Logistic Regression, LDA) may be adversely affected.
- Tree based models and ensemble methods are more robust to outliers and will be targeted as part of the project.



Normality

The assumption of multivariate normality is violated. Predictors show high skewness, heavy tails, and multimodality. These characteristics limit the applicability of parametric models such as LDA, and favour non-parametric models such as tree based and ensemble methods, kNN and SVM may better capture nonlinear relationships, though kNN is sensitive to high dimensionality.

Classification Implementation Plan

We have completed data cleaning & wrangling, as well as EDA.

Next steps are as follows.

(show in visual: For high dimensional data, can you plot the data in a lower dimension for visualization purpose (e.g. PCA plots))

Evaluation Metrics

Given the balanced dataset, **accuracy** will serve as our primary metric. However, to better capture nuances, we will also consider:

- **Log Loss**: Penalizes confident but incorrect predictions.
- **ROC-AUC**: Evaluates overall discriminatory ability across thresholds.

Modelling Approach

Based on the EDA, we will perform the following models in order of suitability:

- **Random forest:** easy to interpret, and tune to prevent overfitting. This will be highly regarded as
- **Trees**: easy to interpret but prone to overfitting, which would mean that we would need to include hyperparameter tuning to reduce overfitting.
- **Logistic Regression**: May remain effective given the large sample size, despite assumption violations.
- **LDA**: Likely underperform due to non-normality and outliers.
- **kNN & SVM**: Better suited for non-linear boundaries but challenged by class overlap and dimensionality. kNN is also computationally expensive for large datasets (must compute distance to all training points) and is sensitive to high dimensions and irrelevant features. Sensitive to feature scaling (distance-based). No model coefficients or interpretability which will has lowered this in terms of suitability as we're hoping to interpret features relations to gain insights from the model.
- **Boosting**: sequentially corrects errors; XGBoost adds regularisation.
- We will Compare models, SHAP values/feature importances for interpretation.

Appendix

Feature Names

Variable.Group	Variable.Name	Variable.Description
Self reported Subject	gender	Gender of self
Self reported Subject	d_age	Difference in age
Self reported Subject	race	Race of self
Self reported Subject	samerace	Whether the two persons have the same race or not.
Partner: Preferences Ratings	race_o	Race of partner
Partner: Preferences Ratings	attractive_o	Rating by partner (about me) at night of event on attractiveness
Partner: Preferences Ratings	sincere_o	Rating by partner (about me) at night of event on sincerity
Partner: Preferences Ratings	intelligence_o	Rating by partner (about me) at night of event on intelligence
Partner: Preferences Ratings	funny_o	Rating by partner (about me) at night of event on being funny
Partner: Preferences Ratings	ambitous_o	Rating by partner (about me) at night of event on being ambitious

Variable.Group	Variable.Name	Variable.Description
Partner: Preferences Ratings	shared_interests_o	Rating by partner (about me) at night of event on shared interest
Subject Self: Self-evaluation	attractive	Rate yourself - attractiveness
Subject Self: Self-evaluation	sincere	Rate yourself - sincerity
Subject Self: Self-evaluation	intelligence	Rate yourself - intelligence
Subject Self: Self-evaluation	funny	Rate yourself - being funny
Subject Self: Self-evaluation	ambition	Rate yourself - ambition
Subject Self: Degree of Interest on common Interests	sports	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	tvsports	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	exercise	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	dining	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	museums	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	art	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	hiking	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	gaming	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	clubbing	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	reading	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	tv	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	theater	Your own interests [1-10]

Variable.Group	Variable.Name	Variable.Description
Subject Self: Degree of Interest on common Interests	movies	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	concerts	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	music	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	shopping	Your own interests [1-10]
Subject Self: Degree of Interest on common Interests	yoga	Your own interests [1-10]
Results	decision_o	Decision of partner at night of event.

References

- Fisman, R., Iyengar, S. S., Kamenica, E., & Simonson, I. (2006). Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics, 121*(2), 673–697. https://doi.org/10.1162/qjec.2006.121.2.673
- OpenML. (n.d.). *Speed Dating Dataset (ID 40536)*. Retrieved August 30, 2025, from https://www.openml.org/search?type=data&sort=version&status=any&order=asc&exact_name=SpeedDating&id=40536