Group 51 Progress Report:
The Optimal Match Model: Predicting Ideal Partner with ML

Alvin Qian , Om Patel, Gregory Archer
{qiana2,patelo11,archeg1}@mcmaster.ca

1 Introduction

Online dating platforms often optimize for engagement metrics such as clicks, swipes, or stated preferences rather than genuine mutual compatibility. Many systems therefore prioritize popularity or activity rather than understanding the bidirectional nature of attraction. In this project, we study the problem of predicting an individual's ideal partner profile from structured demographic, preference, and rating data. Our goal is not only to identify who a person is likely to say "yes" to but also to infer what kind of person would, in turn, find them compatible.

Given a participant A, our model predicts the set of individuals A would likely respond positively to, based on their recorded preferences and historical choices. From the top-k candidates, we generate a composite partner profile B^* representing the aggregated traits of A's most desired matches. The model then predicts who B^* would be most likely to say "yes" to, forming another representative profile C^*. Comparing A and C^*, referred to as the preference gap, highlights asymmetries between what an individual seeks and what those they desire tend to prefer. This framing offers insight into unreciprocated attraction and the dynamics of compatibility, while providing both recommendation value and interpretability.

The task is formulated as a binary classification problem (predicting the "yes" or "no" decision) with ranking-based evaluation to measure recommendation quality. Our objectives are threefold:

1. Build a leakage-free, explainable pipeline that predicts an individual's decision outcome.
2. Construct the composite profiles B^* and C^* for analyzing preference gaps.
3. Evaluate performance using metrics such as Accuracy, F1 score, ROC-AUC, and Precision@k.

2 Related Work

Our work is informed by several research areas. The first is the original analysis of our dataset, the Columbia Speed Dating Experiment, where Fisman and Iyengar found that attractiveness, fun, and shared interests were highly predictive of matching decisions. Second, our A→B→C loop concept is inspired by two-sided matching theory (Roth, n.d.), most famously represented by the Gale-Shapley algorithm, which underscores that stable matches must account for the preferences of both sides. Third, our approach is related to the broader field of recommender systems, which often use collaborative or content-based filtering to predict preferences, though typically not in a two-hop "preference gap" framework. Fourth, we draw from psychological research on partner preferences, such as the meta-analysis by Eastwick et al. (2014), which reviewed the predictive validity of stated ideal partner preferences. Finally, given our goal of providing insights, our work connects to explainable AI (XAI) methods, such as SHAP, which we plan to use to identify which features most influence compatibility predictions.

3 Dataset

We are using the Columbia Speed Dating dataset, as described in our proposal. It consists of **8,378 observations** (individual dates) and **123 columns**.

3.1 Preprocessing and Cleaning

The raw dataset required significant preprocessing. Our pipeline performs the following steps:

- **String Decoding:** Many text columns were encoded as Python byte literals (e.g., 'b'female''). We decoded these into standard UTF-8 strings.
- **Normalization:** All string values were converted to lowercase and stripped of lead-

ing/trailing whitespace to ensure consistency (e.g., "Law" and "law" are treated as identical).

- **Missing Values:** We unified various missing value markers (e.g., "na", "n/a", "", "nan") into a single ‘pd.NA‘ representation.

- **Numeric Coercion:** Columns that appeared to be numeric but were stored as objects (e.g., "1.0", "5") were automatically coerced into floating-point types, while preserving categorical ranges (e.g., "[0-1]").

This stable ID generation is crucial for our participant-based train-test split.

4 Features

The model inputs were derived exclusively from pre-event survey responses in the Speed Dating dataset to ensure fairness and avoid post-event leakage. Each training example represents a pairing between two participants (A and B), with features combining demographic, self-assessment, and preference information from both sides. Using only pre-date data ensures that the model learns compatibility patterns rather than reactions or biases formed during the actual event.

For participant A, features included demographic attributes (age, gender, race, and field of study), self-ratings (attractive, sincere, intelligence, funny, ambition), personal interests (e.g., sports, music, movies, reading, exercise, hiking, art, shopping), and stated partner preferences (attractive_important, sincere_important, etc.). For participant B, corresponding partner features were used, including demographic information (age_o, gender_o, race_o), self-ratings (attractive_o, sincere_o, intelligence_o, funny_o, ambitious_o), and stated partner preferences (pref_o_attractive, pref_o_sincere, etc.). This alignment allowed the model to learn the relationship between what A looks for in a partner and how B describes themselves, creating a symmetric and interpretable feature space.

Feature engineering included converting all byte-encoded categorical fields to strings, coercing numeric columns to float representations, and one-hot encoding categorical attributes. Derived attributes such as age difference and same-race indicators were added to capture potential compatibility signals. Missing demographic or preference values were imputed using the mode for categorical features and the median for numerical ones to maintain data consistency without distorting distributional properties. All numerical features were normalized to a [0,1] range to ensure balanced model learning, particularly for tree-based methods that may otherwise overweight higher-magnitude attributes.

No explicit dimensionality reduction or learned embeddings were applied at this stage, as the dataset's moderate size made full feature inclusion tractable. However, future iterations may experiment with principal component analysis (PCA) or learned latent embeddings to capture higher-order interactions between traits. We also plan to use model-based feature importance scores (e.g., from XGBoost or SHAP values) to identify redundant or weakly predictive inputs and better interpret the contribution of personality versus demographic variables to match likelihood. These interpretive tools will guide future pruning and feature weighting decisions, improving both transparency and model efficiency.

5 Implementation

We implemented a binary classification model to predict whether a participant (A) would say “yes” to another participant (B) using pre-event survey data from the Speed Dating dataset. The target label is the `decision` column, where 1 indicates “yes” and 0 indicates “no.” The modeling process was designed to ensure reproducibility, fairness, and the avoidance of post-event data leakage.

Our baseline model is a simple majority class predictor that always predicts “no,” achieving approximately 58% accuracy. This baseline corresponds to the dataset's natural class imbalance (58/42 split) and serves as a lower bound for meaningful model comparison.

For our primary model, we used a calibrated XGBoost classifier, a gradient-boosted decision tree model known for its robustness to heterogeneous data types and ability to capture nonlinear feature interactions. The model was trained using binary cross-entropy (log loss) as the objective function, optimized via additive tree boosting. We performed hyperparameter tuning using randomized search over parameters such as `max_depth`, `learning_rate`, `subsample`, and `n_estimators`, balancing predictive performance and overfitting risk. Early stopping was used with a validation patience of 50 rounds to halt training when no further

improvement was observed.

To improve probability calibration, we applied isotonic regression on validation outputs so that predicted probabilities aligned more closely with empirical decision frequencies. This calibration step was important for interpretability, as our later analyses depend on comparing predicted likelihoods across simulated participants rather than binary outputs alone.

The input features included both A's and B's pre-event survey responses: demographic variables (age, gender, race, and field of study), self-assessments (attractiveness, sincerity, intelligence, humor, ambition), personal interests, and partner preference indicators (`_important` and `pref_o_`). Post-event attributes such as `like`, `match`, and `decision_o` were excluded to prevent label leakage.

Data splits were performed at the participant level rather than by row, ensuring that information about any one participant did not appear in multiple sets. The data were divided into training (70%), validation (15%), and test (15%) partitions. Model optimization followed XGBoost's gradient-based boosting with learning rate scheduling, and results were tracked using both cross-validation and held-out evaluation.

Current results yield an accuracy of approximately 0.63, F1 score of 0.51, and ROC-AUC of 0.66, representing a clear improvement over the 0.58 baseline. Feature importance analysis shows that self-rated attractiveness, shared interests, and age difference are among the most influential predictors of a positive decision.

Implementation challenges primarily involved dataset cleaning and ensuring interpretability. Many categorical fields were stored as byte strings requiring decoding and normalization, and balancing explainability with model complexity remained an ongoing consideration. Future iterations will extend this framework into the A $\rightarrow$ B $\rightarrow$ C pipeline to simulate multi-stage preference dynamics and analyze how compatibility signals propagate across indirect match predictions.

6 Results and Evaluation

Model evaluation is ongoing, with current experiments focused on establishing baseline performance and validating the modeling pipeline. All results reported here are preliminary and based on the current version of the calibrated XGBoost classifier.

The dataset was divided into 70% training, 15% validation, and 15% testing sets, split at the participant level to ensure that no individual appeared in multiple subsets. This setup prevents information leakage and provides an unbiased framework for model tuning and evaluation.

We trained an XGBoost classifier using binary cross-entropy loss and applied isotonic calibration on validation predictions to improve probability reliability. Calibration quality was assessed using reliability curves and the Brier score, which provided early indications that the model's predicted probabilities align reasonably well with observed outcomes.

Evaluation focuses on standard binary classification metrics, including accuracy, precision, recall, F1 score, and ROC-AUC. These metrics are being used to assess both discrimination ability and class balance. Preliminary results show an accuracy around 0.63, F1 score near 0.51, and ROC-AUC of approximately 0.66, which represents a moderate improvement over the 0.58 baseline from the majority-class predictor. While these numbers are expected to evolve with further tuning and feature refinement, they suggest that pre-event attributes carry meaningful predictive information.

Ongoing work involves refining hyperparameters, experimenting with alternative model architectures, and expanding evaluation to include ranking-based metrics such as Precision@k. We also plan to incorporate SHAP-based feature importance analysis to better understand which demographic and personality factors most strongly influence predicted compatibility scores. Future iterations will integrate these insights into the A $\rightarrow$ B $\rightarrow$ C pipeline to evaluate how prediction quality scales in multi-stage matching.

7 Feedback and Plans

The feedback provided by our TA emphasized three key areas for improvement: establishing a clear baseline model for comparison, expanding evaluation metrics beyond accuracy and F1 score, and including more experimental details such as hyperparameters, visualizations, and training summaries. These points align well with our current development roadmap and will guide our next phase of work.

First, we plan to formalize the baseline by implementing and reporting results for multiple simple

models, such as logistic regression and decision trees, in addition to the majority-class predictor. This will help contextualize our XGBoost model's performance and quantify improvement more concretely. We will also include baseline metrics in a results table for clarity.

Second, we aim to broaden our evaluation by incorporating additional performance metrics. In particular, we plan to include Precision@k and Mean Reciprocal Rank (MRR) to better capture the model's ranking capability. Potentially the Brier score and reliability curves to evaluate calibration quality as well. These metrics will provide a more nuanced understanding of both predictive accuracy and probabilistic reliability, which are important for the planned A $\rightarrow$ B $\rightarrow$ C compatibility framework.

Third, we will expand the experimental section to include visual and quantitative summaries of model behavior. Planned additions include feature importance plots from XGBoost and SHAP, learning curves showing convergence trends, and calibration plots comparing predicted and observed probabilities. These visualizations will make our analysis more interpretable and strengthen the report's empirical depth.

Finally, we recognize opportunities to refine the implementation itself. We intend to conduct more extensive hyperparameter tuning using randomized or Bayesian optimization, experiment with additional ensemble models (e.g., LightGBM or CatBoost), and evaluate whether dimensionality reduction or latent embeddings improve performance. We also plan to explore balancing techniques such as class weighting or SMOTE to mitigate the dataset's inherent imbalance.

Overall, the TA's feedback provided clear and actionable directions. Our remaining work will focus on strengthening the experimental rigor, improving interpretability, and extending the analysis to the full A $\rightarrow$ B $\rightarrow$ C pipeline. These steps will help ensure that our final model is both robust and explainable, providing a more comprehensive understanding of human compatibility prediction.

Team Contributions

Om led the model development and implementation process, including data preprocessing, feature engineering, and training the XGBoost classifier. He also integrated the calibration and evaluation framework and coordinated the A $\rightarrow$ B $\rightarrow$ C pipeline design.

Gregory focused on exploratory data analysis, dataset cleaning, and the construction of participant-level profiles. He was responsible for identifying key features from the survey data and preparing visualizations and summary statistics used in both the report and presentation materials.

Alvin contributed to the experimental design and evaluation setup, helping define baseline comparisons and select appropriate performance metrics. He also supported documentation, result interpretation, and report writing, ensuring clarity and consistency across all sections.

References

Paul W. Eastwick, Laura B. Luchies, Eli J. Finkel, and Lucy L. Hunt. 2014. The predictive validity of ideal partner preferences: a review and meta-analysis. *Psychological Bulletin*, 140(3):623–665. Epub 2013 Apr 15. PMID: 23586697.

Alvin E. Roth. n.d. Matching (two-sided models). Retrieved November 11, 2025.

Tables and Figures

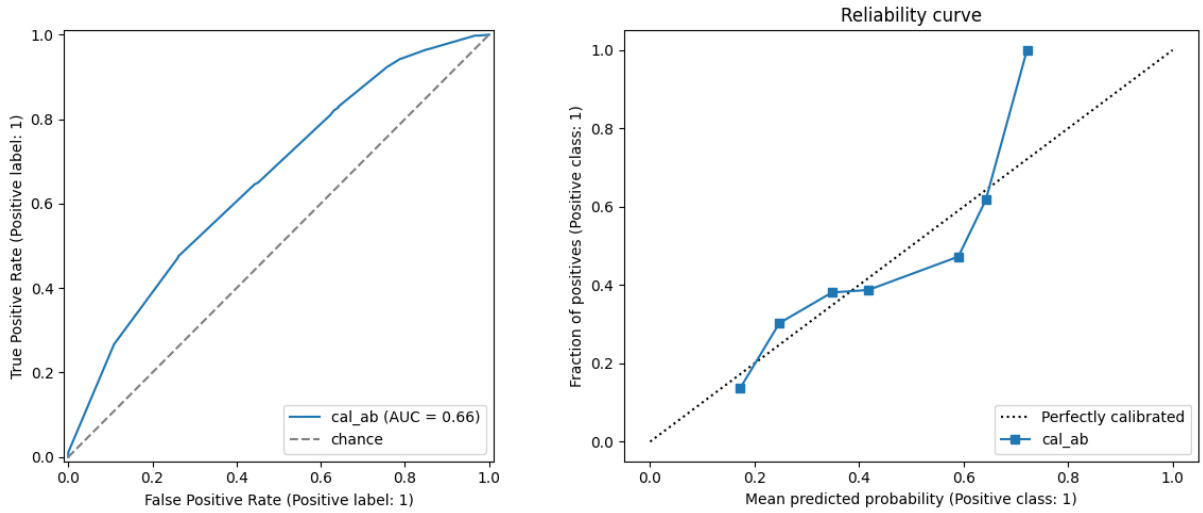

Figure 1: Distribution of predicted probabilities by actual outcome, showing the model's ability to discriminate between true positives and true negatives.

ID	Gen.	Age	Race	Field	Att.	Sin.	Int.	Fun.	Amb.	Spo.	Mus.	Mov.	Rea.	Exe.	p_AB
137	F	29	Oth.	Music Ed.	9	10	9	9	8	8	10	9	5	10	0.64
439	F	27	Eur.	Finance	8	10	10	9	10	7	10	10	5	10	0.64
367	F	26	Lat.	Law	9	9	9	9	9	8	9	9	7	7	0.64
199	F	29	Eur.	Psychology	7	8	4	8	8	6	6	7	7	4	0.64
198	F	28	Eur.	Social Work	9	8	5	9	3	6	6	9	9	5	0.64
369	F	28	Eur.	German Lit.	7	10	7	10	7	1	10	10	10	5	0.64
370	F	29	Oth.	Psychology	7	9	9	9	9	3	9	7	5	9	0.64
194	F	22	Eur.	Social Work	8	9	7	10	7	8	5	7	9	5	0.64
382	F	22	Eur.	Comm.	7	9	9	9	4	2	10	10	4	1	0.64
383	F	22	Eur.	Social Work	6	8	8	8	8	7	10	8	6	10	0.64

Table 1: Top 10 recommended partners for sample participant A.

Attribute	Value
Age	26.2
Attractive	7.7
Sincere	9.0
Intelligence	7.7
Funny	9.0
Ambition	7.3
Sports	5.6
Music	8.5
Movies	8.6
Reading	6.7
Exercise	6.6
Gender	Female
Race	European/Caucasian-American
Field	Social Work

Table 2: Composite profile B* derived by averaging the top 10 recommended partners for participant A. Numeric values represent means across all candidates; categorical values represent the mode (most common value).