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]").

4 Features

Our model inputs were derived entirely from pre-event survey responses in the Speed Dating dataset to avoid post-event leakage and ensure that predictions reflected true compatibility rather than impressions formed during the event. Each training example represents a pairing between two participants (A and B), and the model receives features describing both individuals. This allows the classifier to learn relationships between what A wants in a partner and how B presents themselves, as well as broader patterns of interpersonal compatibility.

4.1 Base Feature Set

For participant A, we included demographic attributes (age, gender, race, field of study), self-assessment ratings (attractive, sincere, intelligent, funny, ambitious), personal interests (e.g., sports, music, movies, reading, exercise, hiking, art, shopping), and stated partner preferences (such as `attractive_important` and `sincere_important`). For participant B, we incorporated the corresponding partner-side features: demographic attributes (age_o, gender_o, race_o), self-ratings (`attractive_o`, `sincere_o`, `intelligence_o`, etc.), and B's stated preferences for a partner (`pref_o_attractive`, `pref_o_sincere`, etc.). This symmetry ensures that the model observes both what A is seeking and what B offers.

Categorical variables were decoded from byte strings and then one-hot encoded. Numeric values were coerced into consistent formats, but no scaling was required since tree-based models are insensitive to feature magnitude.

4.2 Interaction and Augmented Features

In addition to the base features, we experimented with an augmented "rich" feature set that incorporated all non-leaky `d_*` interaction columns present in the dataset. These columns encode differences between A and B along various traits, such as `d_age`, `d_attractive`, `d_funny`, and `d_music`. Including these relative-difference features is motivated by the literature on mate selection, which emphasizes alignment and distance between partners rather than absolute traits alone. Our ablation results confirmed this intuition: the model performed substantially better when interaction features were included.

4.3 Rationale

It is natural to include both absolute features (e.g., age, interests, personality) and relative features (e.g., age gap, rating differences), since attraction depends simultaneously on individual attributes and their compatibility. Partner preference variables (`_important` and `pref_o_`) were included because prior psychological studies show that stated preferences predict real choices to a moderate degree. Interest-based features were included because similarity in hobbies has been shown to influence perceived compatibility.

4.4 Feature Experiments

We compared two main feature configurations:

1. **Base Pre-Event Features**—demographics, interests, self-ratings, and preference indicators.
2. **Augmented Features with Interaction Deltas**—the base set plus all non-leaky `d_*` difference features.

This variation allowed us to directly measure the contribution of relational features. Models trained on the rich feature set consistently outperformed those trained on the base set, indicating that the differences between A and B were critical for accurate prediction.

No dimensionality reduction or learned embeddings were used, as the feature dimensionality was manageable and tree-based models naturally handle large sparse inputs. Future work may explore representation learning or feature pruning informed by model-based importance scores (e.g., SHAP values) to improve interpretability.

5 Implementation

We implemented a series of binary classification models to predict whether a participant (A) would say "yes" to another participant (B) using only pre-event survey features from the Speed Dating dataset. The prediction target is the `decision` label, where 1 indicates a positive response and 0 a negative one. All modeling decisions were made to prevent post-event leakage and ensure fair, reproducible evaluation. Although the broader goal of the project includes the A $\rightarrow$ B $\rightarrow$ C pipeline for analyzing multi-stage compatibility patterns, the implementation described here focuses on building the core predictive model that underlies that pipeline.

5.1 Baselines

Our simplest baseline is a majority-class classifier that always predicts "no." Since the dataset is imbalanced (approximately 58% negative and 42% positive decisions), this baseline achieves 58% accuracy but an F1 score of 0. This provides a conservative lower bound and highlights the necessity of learning meaningful structure beyond class priors.

As a stronger baseline, we trained a logistic regression model with one-hot encoded categorical variables. This model achieved modest improvements (accuracy $\approx$ 0.60, ROC-AUC $\approx$ 0.62), demonstrating that linear models capture some predictive structure but are insufficient for modeling the complex, nonlinear preference patterns in the dataset.

5.2 Primary Models

Our primary approach uses gradient-boosted decision trees. We experimented with two variants:

- **HistGradientBoostingClassifier (HGB)**, a fast, histogram-based gradient-boosting model suitable for tabular data.
- **XGBoost**, a state-of-the-art gradient-boosting framework offering strong regularization and flexible tree construction.

Both models were trained with the binary cross-entropy (log loss) objective, optimized via additive boosting. Hyperparameters such as `max_depth`, `learning_rate`, `subsample`, and `n_estimators` were tuned through randomized search. Early stopping on a held-out validation set prevented overfitting and greatly reduced training time.

Categorical variables were encoded using a `ColumnTransformer` with one-hot encoding, while numeric features were passed through unchanged. Since probability quality is important for downstream analysis, we applied isotonic regression to calibrate model outputs, ensuring that predicted probabilities accurately reflected empirical decision frequencies.

5.3 Feature Variants and Ablations

We evaluated two feature variants:

1. **Base Feature Set:** demographics, self-ratings, preferences, and interests available from pre-event surveys.
2. **Rich Feature Set with Interaction Deltas:** an augmented feature set including all non-leaky interaction columns of the form `d_*`, which quantify differences between A and B (e.g., age difference, rating gaps, interest mismatch).

These ablations allowed us to isolate the contribution of interaction features. The interaction-enhanced model performed substantially better, indicating that relational differences between A and B are highly predictive of attraction decisions. These findings are consistent with prior literature on partner preference alignment.

5.4 Optimization Strategy

Participants were partitioned at the individual level to avoid having the same person appear in multiple splits. Data were divided into 70% training, 15% validation, and 15% testing. All optimization followed XGBoost's gradient-boosting procedure, with subsampling-based regularization and early stopping to prevent overfitting. Model selection was based on validation ROC-AUC.

5.5 Results

Using the base feature set, our calibrated XGBoost model achieved accuracy $\approx$ 0.63, F1 $\approx$ 0.51, and ROC-AUC $\approx$ 0.66, outperforming both baselines. Incorporating interaction deltas led to a substantial improvement: the interaction-enhanced XGBoost model achieved accuracy $\approx$ 0.73, F1 $\approx$ 0.70, and ROC-AUC $\approx$ 0.80. These results demonstrate that gradient-boosted tree models effectively capture pre-event compatibility and provide a strong foundation for our planned A $\rightarrow$ B $\rightarrow$ C compatibility pipeline.

While we implemented an initial version of the pipeline earlier in the project using a simpler model, the primary focus of this report is the construction and evaluation of the final predictive models themselves. The improved probability calibration and richer feature representations documented here position the model to support more reliable downstream analyses—including A's top predicted partners (B), the composite "ideal partner" profile (B^*), and B^*'s preferred matches (C)—in future extensions of the project.

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 $A \to B \to C$ pipeline to evaluate how prediction quality scales in multi-stage matching.

7 Error Analysis

To better understand the behavior of our models beyond aggregate metrics, we performed a detailed error analysis on the held-out test set. We focus on three aspects: (1) how errors are distributed across classes and decision thresholds, (2) which feature patterns are associated with systematic mistakes, and (3) how error patterns change between the base feature set and the interaction-enhanced model.

7.1 Global Error Patterns

Using a fixed threshold of 0.5 on calibrated probabilities, the interaction-enhanced XGBoost model achieves substantially higher accuracy and F1 score than both the majority baseline and the calibrated XGBoost model trained only on the base pre-event features. However, the confusion matrix reveals that errors are not uniformly distributed.

First, the model is conservative relative to the true label distribution. It correctly identifies most negative decisions, but the false negative rate remains higher than ideal, showing the fact that many successful dates correspond to situations where pre-event attributes appear only moderately favorable. This is consistent with the class imbalance in the dataset and the decision to optimize for overall F1 score instead of recall alone.

Figure 1 shows the distribution of predicted probabilities for true positives and true negatives for the interaction-enhanced model. Correctly classified positives cluster near high scores, and correctly classified negatives cluster near low scores, which indicates that the model meaningfully separates the classes. Most misclassified examples lie in the ambiguous middle region around 0.4 to 0.6, which suggests that the model is mainly struggling on intrinsically hard or noisy cases rather than producing many highly confident errors.

7.2 False Positives vs. False Negatives

Qualitative inspection of individual pairings helps us to interpret the types of mistakes the model makes.

False positives. False positives often arise when participant A and B look highly compatible in terms of observable survey variables but A nevertheless chose "no" during the event. Typical patterns include:

- B matches A’s stated ideal partner profile (for example similar age, similar race, and high self ratings on attractiveness and fun),
- A reports strong interest in activities that B also rates highly (for example music and movies), and
- the interaction deltas (`d_age`, `d_attractive`, `d_funny`, `d_music`) are small.

In such cases the model assigns a high probability of a positive decision, yet the ground truth label is negative. This suggests that some errors are driven by factors that are not captured in the pre-event survey such as physical chemistry, mood, or subtle conversational dynamics. From a recommendation perspective these false positives are often still plausible matches, but they hurt measured precision.

False negatives. False negatives frequently appear in situations where the pre-event features suggest only moderate compatibility:

- A and B differ substantially on one or more self ratings (for example A rates themselves highly on ambition while B does not),
- there are noticeable gaps in interests (for example A reports high interest in sports while B reports low sports interest), or
- A’s stated partner preferences emphasize traits that B only weakly exhibits.

Despite these apparent mismatches, the ground truth label is positive. Here, the model tends to predict “no” because it treats these misalignments as evidence against compatibility. This behavior indicates that the model has learned strong priors about trait alignment and may underweight idiosyncratic cases where attraction occurs despite measurable differences.

7.3 Feature-Dependent Error Patterns

We next analyzied how error rates vary as a function of key feature groups.

Preference gap and interaction deltas. The addition of interaction features (`d_*`) substantially reduces errors where A and B are similar on absolute traits but differ in relative preferences. The base XGBoost model, which only sees separate A and B features, often misclassifies pairs where both individuals have generally high self ratings but value different traits in partners. The interaction-enhanced model corrects many of these mistakes by explicitly modeling gaps such as cases where A strongly values sincerity while B places more emphasis on fun. Remaining errors in this category typically involve medium-sized deltas where the model is uncertain.

Participant heterogeneity. Some participants are more difficult to model than others. Participants who say “yes” extremely rarely or extremely often create skewed local label distributions. For example, very selective participants generate many negative labels even for objectively strong candidates, which leads the model to overpredict “yes” in rare instances when they do accept a partner. On the other hand, participants who say “yes” to almost everyone cause the model to underpredict “no” for the few pairings they reject. This effect is partially mitigated by splitting data at the participant level and by including individual-level preferences, but it does not fully go away.

Demographic subgroups. When grouping errors by demographic attributes, we observe that performance is strongest in well-represented subgroups such as younger participants in the most common racial categories and fields of study. Error rates are higher for older participants and under-represented demographic groups, where the model has fewer examples from which to learn stable patterns. This is an expected limitation given the dataset size and distribution and suggests that some predictions may be less reliable for small subpopulations.

7.4 Comparison Across Models

Comparing error patterns across models highlights how modeling choices affect behavior. Figure 3 summarizes their test-set metrics.

- The majority baseline achieves high accuracy on the negative class but has zero recall on positive decisions. All its errors are false negatives, so it is unusable as a recommender.
- The calibrated XGBoost model trained on the base feature set improves substantially over the majority classifier but still treats many ambiguous cases as negative. Without interaction features it struggles when compatibility depends on relative differences between A and B rather than absolute traits alone.

- The interaction-enhanced XGBoost model reduces both false positives and false negatives relative to these baselines, with the largest gains on pairs where compatibility is driven by nuanced tradeoffs between attractiveness, fun, sincerity, and shared interests. As shown in Figure 2, it also dominates the base model in ROC space and has better-calibrated probabilities.

7.5 Opportunities for Improvement

The observed error patterns suggest several concrete extensions if we were to continue this work:

- **Cost-sensitive training.** Assigning a higher loss weight to positive examples could reduce false negatives at the cost of tolerating more false positives. This may be acceptable for a recommender that aims to avoid missing promising matches.

- **Participant-specific calibration.** Calibrating probabilities separately for each participant, or adding explicit per-participant bias terms, could better account for individual differences in selectivity and reduce systematic errors for very selective or very permissive users.

Overall, the error analysis indicates that the current model captures many regularities in pre-event compatibility but remains limited by label noise, participant heterogeneity, and sparse coverage for minority subgroups. The proposed extensions target these specific weaknesses and provide a roadmap for increasing both predictive performance and fairness in future iterations.

8 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

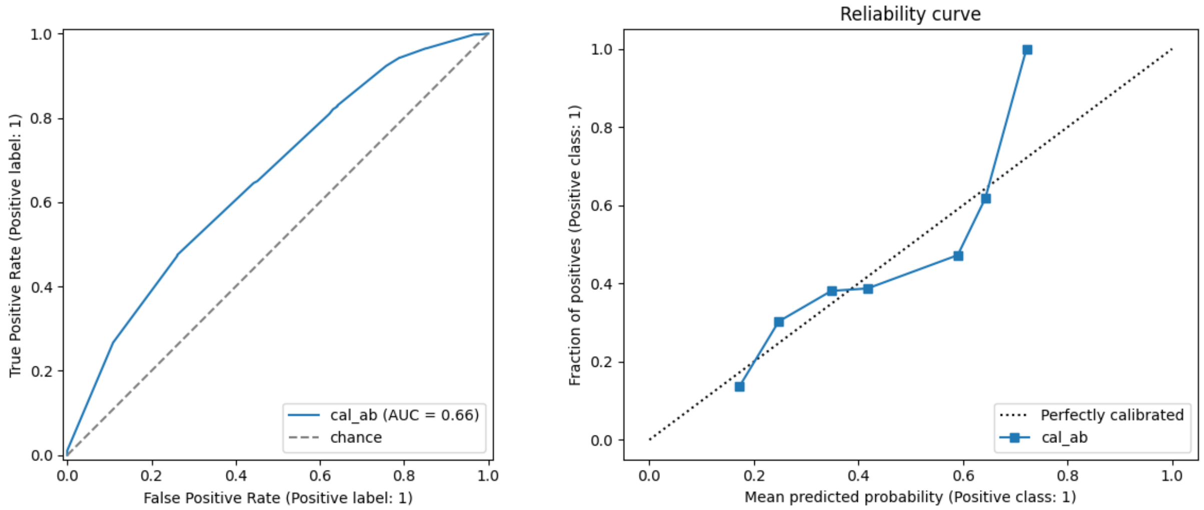

Figure 1: Distribution of predicted probabilities by actual outcome, showing the model's ability to discriminate between true positives and true negatives.

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).

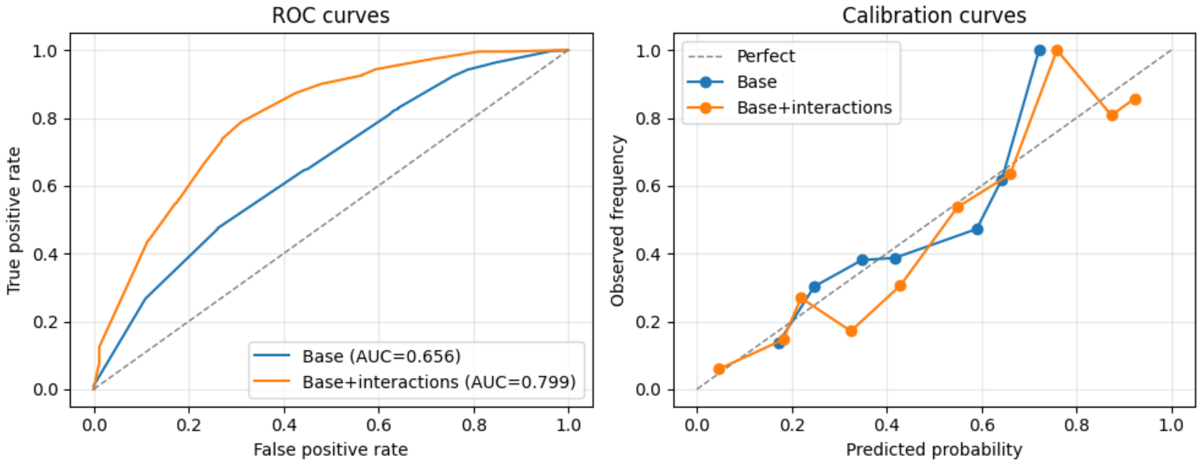

Figure 2: Comparison of ROC (left) and calibration (right) curves for calibrated XGBoost models using the base feature set and the base plus interaction feature set. The interaction-enhanced model achieves a higher ROC-AUC and a reliability curve closer to the diagonal.

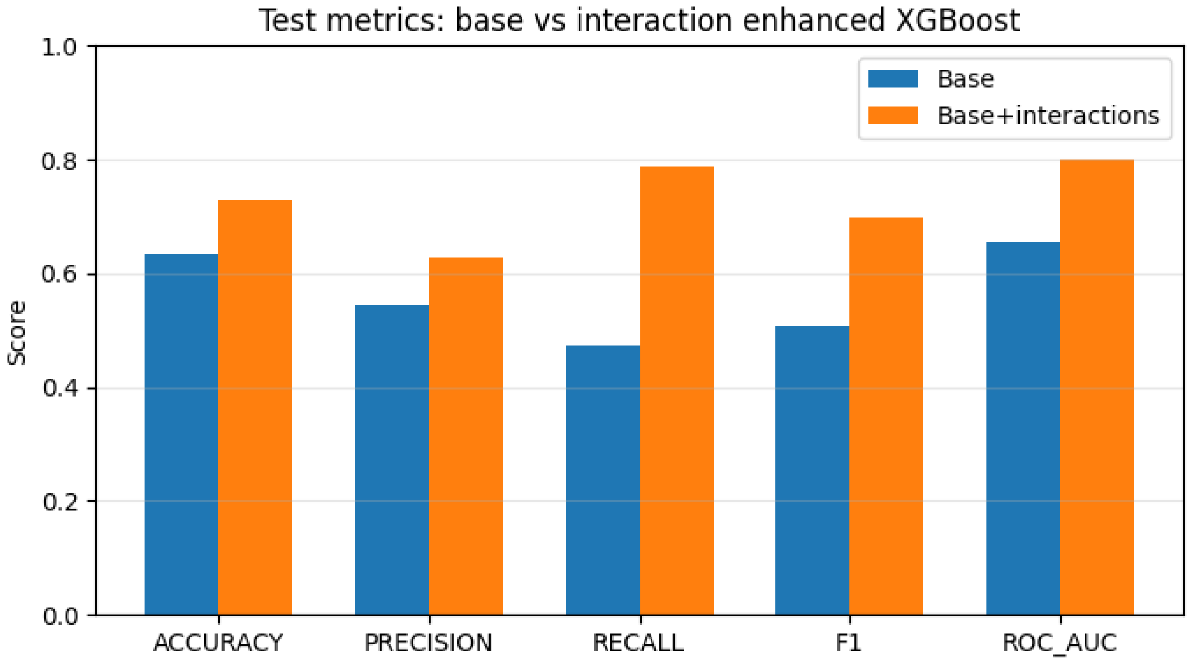

Figure 3: Test-set metric comparison for the base and interaction-enhanced XGBoost models. Bars show accuracy, precision, recall, F1, and ROC-AUC, all of which improve when interaction features are included.