PREDICTING DATING MATCHES USING MACHINE LEARNING

MENG ZHANG

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2113115

COMMITTEE

dr. Seyed Mostafa Kia
drs. Mohammad Zamanzadeh Nasrabadi

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 20th, 2023

WORD COUNT

7463

ACKNOWLEDGMENTS

I would like to thank my supervisor Kia, I can't make this thesis without him. And his course Advanced Data Processing is one of my favourite course in this program, as it make me realize the importance to read, understand and learn from the documents, which helps me a lot when I code and want to build something by myself. And I would like to thank my family support me to resign my job, go back to school and live abroad. I know this is a hard decision for them but thanks they still do it.

CONTENTS

PREDICTING DATING MATCHES USING MACHINE LEARNING

MENG ZHANG

Abstract

Intimate relationships and marriage are deeply meaningful in people's lives, offering emotional support, personal growth, and fulfillment. This study aims to explore how well machine learning models can predict suitable matches among various dating candidates, potentially enhancing relationship satisfaction and reducing biases that can arise from human judgment. It also seeks to expand the use of machine learning in matchmaking and operations research. Based on the findings of this study, machine learning models like RF, XGBoost, and LightGBM prove effective in predicting dating matches and decisions made by both female and male candidates. LightGBM with random oversampling performs best overall in predicting dating matches. For predicting decisions made by female candidates, LightGBM with class weight shows the highest effectiveness, while XGBoost performs better for male candidates. To improve the understanding of how these models influence dating decision-making, the study utilizes the SHapley Additive exPlanations (SHAP) technique. This approach identifies key factors that influence candidates' dating decisions. For females, income and the fun level of potential partners, along with their own income, are crucial factors. For males, attractiveness of potential partners, their age, and their own income are the most influential factors in making dating choices.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Source/Code/Ethics/Technology Statement Example*

The dataset used in this study comes from OpenML[1], collected by Fisman and Iyengar (Fisman et al., 2006) and involving human participants. The original owner of the data and code used in this thesis retains ownership during and after completion. However, since the dataset is posted on an open-source platform, it can be freely accessed and used.

[1] https://www.openml.org/search?type=data&sort=runs&id=40536&status=active

All figures in this thesis belong to the author. The thesis code can be accessed via the GitHub repository at the following link. These codes are used for reference[2]. Tools such as ChatGPT[3], Grammarly[4], and Reverso[5] are used to improve the author's original content, including paraphrasing, spell checking, and grammar. Overleaf[6] is used for typesetting tools and services.

2 INTRODUCTION

Intimate relationships are profoundly important in individuals' lives, offering emotional support, personal growth, and fulfillment (Howe, 2002). They fulfill our fundamental desires for love and belonging (Muniruzzaman, 2017), enhancing overall well-being and quality of life (DEMETER & PAIUSAN, 2018). The factors influencing the formation of intimate relationships are complex and include attachment styles, personality traits, self-monitoring tendencies, past relationships, dating goals, and substance use (Cruces et al., 2015; Dietrich, 2016; Schindler et al., 2010).

When pursuing romantic relationships, we assess others by understanding their personalities, interests, cultural backgrounds, and experiences to determine compatibility. However, if there were a model that could automatically match individuals and provide stable, optimal results, we might have more time and energy to engage meaningfully with real people.

2.1 *Research Questions*

This research aims to address the dating match problem using machine learning to enhance the performance of dating matches, potentially increasing relationship satisfaction. Furthermore, the findings could broaden the application of machine learning methodologies in the fields of matchmaking and operations research.

> *To what extent can machine learning models predict suitable matches among various dating candidates?*

RQ1 *What is the performance of machine learning models (Random Forest (RF), XGBoost and LightGBM) in predicting dating matches?*

[2] https://www.kaggle.com/code/bextuychiev/model-explainability-with-shap-only-guide-u-need
[3] https://chatgpt.com
[4] https://app.grammarly.com/
[5] https://www.reverso.net/
[6] https://www.overleaf.com/

RQ2 *How well do the models perform when trained separately for males and females?*

RQ3 *Which features are most relevant and important for males and females, respectively, in predicting dating matches based on the model with the best performance?*

2.2 *Relevance*

From a societal perspective, finding the best-performing model directly improves how accurately we can predict dating outcomes. This advancement can lead to better matches between individuals seeking relationships, making it easier to start dating when candidates share similarities or common interests. This can potentially lead to more successful relationships, increased relationship satisfaction, and greater personal happiness. Moreover, using machine learning models can help reduce biases in matchmaking processes, making dating platforms fairer and more inclusive.

From a scientific viewpoint, exploring and comparing different machine learning models for predicting dating compatibility extends their application beyond dating apps. These models can be used in recommendation systems, personalized marketing, and addressing various matchmaking challenges such as college admissions, job placements, and matching patients with suitable doctors. Such research holds significant implications across diverse fields where effective matchmaking is essential.

2.3 *Findings*

Based on the findings of this study, it is revealed that machine learning models such as RF, XGBoost, and LightGBM excel in predicting dating matches and decisions made by both female and male candidates. Specifically, when considering metrics such as F1-score, Matthews Correlation Coefficient (MCC), Balanced Accuracy (BACC), and AUC, along with prediction balance, LightGBM with random oversampling emerges as the top performer for predicting dating matches overall. When predicting decisions made by female candidates, LightGBM with class weight proves most effective, while XGBoost shows better performance in predicting decisions made by male candidates.

To enhance the interpretability of these models in dating decision-making, the study utilizes the SHAP technique. This method helps uncover the key features influencing how candidates make their dating decisions. For female candidates, factors such as the income and fun level of potential partners, along with their own income, emerge as the most crucial.

Meanwhile, for male candidates, factors such as attractiveness level, age of potential partners, and their own income are identified as the top influencer in their dating choices.

3 RELATED WORK

In this section, we will explore related works in the dating match field and techniques to address the issues encountered in this study. Firstly, we will discuss key features that influence people's dating choices and marriage preferences. Next, we will examine the algorithms used by previous researchers to tackle dating matches and other matching problems. Afterward, we will explore techniques for handling imbalanced datasets and methods to measure model performance suitable for such datasets. Following this, we will detail techniques for analyzing feature importance. Finally, we will identify gaps in previous research and explain the methods employed in this study.

3.1 *Key Attributes for Date Selection*

To understand the factors influencing dating selection, research on marriage preferences has been examined. Generally, people's selection behaviors are influenced by attributes such as physical attractiveness (age, appearance, height, weight) and socioeconomic status (intelligence, education, career field, wealth, ambition, kindness). Fisman's work indicates that women prioritize their partners' intelligence, race, and whether they are from an affluent neighborhood, while men emphasize physical attractiveness. Notably, for men, high intelligence and ambition in potential partners (women) become drawbacks when they exceed their own (Fisman et al., 2006). Belot's research indicates that physical attractiveness is important for both genders, but they have different preferences: women prefer younger and taller men, while men are attracted to younger and thinner women. Similar to Fisman's research, Belot finds that people's preferences are also related to economic reasons. They found a relationship between physical attributes (age, height, and weight) and socioeconomic status (education and occupation) (Michèle & Francesconi, 2006). Furnham's research shows that females place more importance on cognitive ability, education, occupation, wealth, and ambitiousness than males, while men value good looks and heredity more than women, aligning with Fisman's findings. However, due to the age range of the candidates, there is no strong preference based on age (Furnham & Tsoi, 2012). Buss's research reviews the science of human mate preferences and their behavioral manifestations. According to Buss's work, males and females employ different long-term mating

strategies. For males, three factors influence their mating selection strategy: identifying potential mates of high reproductive value (relative youth, beauty, and female body attractiveness), paternity certainty (faithfulness and sexual loyalty), and establishing a mutually cooperative and beneficial long-term relationship with some degree of division of labor (kindness, dependability, and good health). For females, three requirements are identified from ancestral women: status and economic resources for themselves and their children, physical protection for themselves and their children, and enhanced mating success of their children. These reasons have evolved into modern preferences: economic resources, ambition, industriousness, and social status; love and kindness; physical formidability, size, athletic prowess, and bravery (Buss & Schmitt, 2019).

3.2 *Algorithms*

Dating match is a classic challenge in operations research's two-sided matching field, often simplified in machine learning as a binary classification problem. Gale and Shapley first established stable matching solutions in 1962, proposing the deferred acceptance algorithm as an efficient method (Gale & Shapley, 1962).

Recently, machine learning and deep learning approaches have been applied extensively to solve matching problems. Mühlbauer and Weber conducted a comparative study on labor market matching, finding that RF outperformed statistical methods by 0.1-1.0 percentage points (Mühlbauer & Weber, 2022). Ravindranath explored deep learning techniques, focusing on trade-offs between strategy-proofness and stability in matching algorithms (Ravindranath et al., 2023). Moreover, beyond linear models, methods like Factorization Machines (FM) and Deep & Cross Network (DCN), which handle feature interactions, have been employed for pairing problems, using metrics such as AUC, log loss, and precision-recall curves (L. Zhang et al., 2019). Alomrani applied reinforcement learning to dynamic matching scenarios, showcasing promising results in online environments (Alomrani et al., 2021).

In addition to demographic factors, Zang investigated visual and text features to objectively quantify attractiveness, employing models like Support Vector Regression (SVR) and Multi-Layer Perceptron (MLP) with correlation coefficients as evaluation metrics (Zang et al., 2017a). Another study by Zang used profile-based, topological, preference, and facial features to predict reciprocal user contacts, with Support Vector Machine (SVM) achieving the highest MCC score (Zang et al., 2017b). Siraj-Ud-Doulah evaluated various models for binary classification tasks, with RF, SVM, and Logistic Regression (LR) emerging as top performers for accuracy across

both binary and multiple classification problems (Siraj-Ud-Doulah & Islam, 2023).

Recommendation algorithms, although slightly divergent from this thesis's focus, are also widely used in recommending highly compatible user matches. Xia applied collaborative filtering-based algorithms to improve dating match recommendations (Xia et al., 2015).

3.3 *Techniques for Imbalanced Dataset*

Training machine learning models on imbalanced data may lead to biases towards the majority class while ignoring the minority class, which can result in poorer performance in predicting the minority class (Singh & Kumar, 2023). Several solutions exist for handling imbalanced data, including sampling methods, cost-sensitive methods, and kernel-based methods (He & Garcia, 2009). In Richmond's study, the Synthetic Minority Oversampling Technique (SMOTE) and the Adaptive Synthetic Sampling Approach (ADASYN) were used with LR, SVM, RF, and XGBoost to predict whether a person is diabetic or not, achieving better performance after applying SMOTE or ADASYN (Danquah, 2020). Sinem and Kemal applied SMOTE, random undersampling, and clustering-based undersampling with XGBoost, LightGBM, and CatBoost to predict the mortality of COVID-19 patients, finding that all models performed best with SMOTE (Keser & Keskin, 2023). In Samuel's study, SMOTE yielded better results with LR compared to random oversampling and random undersampling in predicting whether a woman has ever experienced a terminated pregnancy (Aderoju & Jolayemi, 2019).

3.4 *Performance Measurement*

In addition to algorithms, the methods used to measure performance are important for evaluating and selecting the best models. Flach discussed various performance evaluation techniques in machine learning, highlighting metrics such as accuracy, true/false positive rates, precision, recall, F-score, AUC-ROC curve, and Brier score (Flach, 2019). Canbek reviewed binary classification metrics, identifying Matthews Correlation Coefficient (MCC) as the most robust, followed by Balanced Accuracy (BACC), Informedness (INFORM), Cohen's Kappa (CK), and Markedness (MARK), with metrics like True Negative Rate (TNR), True Positive Rate (TPR), Negative Predictive Value (NPV), and Positive Predictive Value (PPV) considered less robust (Canbek et al., 2022). Chicco's research emphasized the superiority of MCC over F1 score and accuracy in handling imbalanced datasets, as MCC considers the balance between positive and negative elements in

the confusion matrix, ensuring more reliable predictions across all categories (Chicco & Jurman, 2020). Further comparisons of MCC with CK and Brier Score in binary classification reaffirmed MCC's reliability and informativeness, especially in imbalanced datasets (Chicco et al., 2021).

3.5 *Techniques for Feature Importance Analysis*

There are various techniques available to explain the outcomes or behavior of models. Some techniques are intrinsic, such as regression, K-NN, or decision trees, while others are post-hoc. Post-hoc techniques include model-agnostic approaches like partial dependency plots, permutation feature importance, SHAP, Local Interpretable Model-Agnostic Explanations (LIME), Global surrogates, and counterfactual explanations. There are also model-specific post-hoc techniques, such as purity for Random Forests and fuzzy cognitive maps for neural networks (Molnar, 2020).

Among these techniques, LIME and SHAP have been widely used in previous research. LIME is a prominent framework for interpreting machine learning predictions. LIME assumes that all models behave linearly on a local scale, fitting a simple model around a single observation to mimic the behavior of the complex global model in that local area. It uses this local model to explain predictions made by more complex models (Rao et al., 2022).

On the other hand, SHAP provides explanations by computing each feature's contribution to a prediction. Lundberg proposed SHAP values as a unified measure of feature importance due to their ability to provide accurate estimates with fewer evaluations of the original model, thereby enhancing computational efficiency. SHAP's results are consistent with human intuition and can effectively explain differences between classes (Lundberg & Lee, 2017b). Researchers have applied LIME to problems such as text classification using Random Forests and image classification using neural networks, demonstrating its effectiveness in explaining predictions (Ribeiro et al., 2016). Similarly, tree-based SHAP has been applied to Random Forests, gradient boosting, and XGBoost models, showcasing its robust explanation capabilities across different types of ensemble methods (Bifarin, 2023).

3.6 *Research Gaps and Opportunities*

These previous research studies demonstrate the application of machine learning techniques to solve two-sided matching problems in various domains such as labor markets and online matching challenges. However, machine learning models are still relatively new methods for solving dating

match problems. Therefore, based on the findings of previous research, I would like to explore whether these methods can also be applied to address dating match problems.

4 METHOD

4.1 *Dataset Description*

The dataset used in this study was collected from experimental speed dating events spanning the years 2002 to 2004, assembled by professors Fisman and Iyengar for their paper (Fisman et al., 2006). During these events, participants engaged in a series of four-minute 'first dates' with each opposite-sexed attendee. After each conversation, participants were asked to indicate whether they had an interest in seeing their date again and provide ratings for the other attendees on six scales: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. In addition to these ratings, the dataset includes an amount of questionnaire data collected from participants at various stages of the speed dating process. These questionnaires contain demographic information, dating habits, self-perceptions across essential attributes, beliefs, desirable mate qualities, and lifestyle details. Accessible for download via the OpenML platform. The dataset comprises an extensive array of 123 features and encompasses a total of 8,378 instances, offering a comprehensive resource for conducting in-depth analyses into the dynamics of speed dating interactions.

4.2 *Data Cleaning and Processing*

Because the dataset was collected for a slightly different scenario than this thesis, there are two phases for data cleaning and processing, conducted before and after Exploratory Data Analysis (EDA).

Specifically, the original dataset includes data from various stages: registration details (such as candidates' personality, demographic information, event expectations, and date requirements), a survey conducted immediately after the event (indicating matches and date impressions), and two follow-up surveys, one the day after the speed dating event and another 3-4 weeks later. For this thesis, only variables from the registration data and match outcomes are considered, focusing on researching matches based on potential dates' profiles. In the initial data cleaning phase, we removed variables collected after the speed dating event. Additionally, we corrected data types where float values were mistakenly stored as strings and converted all boolean data types to integers. As a final step in this phase, missing values in 'field_cd' and 'career_c' were assigned '-1'.

During the EDA phase, duplicates are removed, as there are 551 candidates but over 8000 records due to each candidate having multiple records for each partner they met.

Based on EDA findings, further data cleaning is performed. Outliers in the gaming and reading features are initially removed.

Next, missing values are addressed by dividing the dataset into two parts: one containing cross data ('iid', 'pid', 'match', 'same race', 'imprace', etc.) with 8368 records, and the other containing self-data with 551 records. Records with missing 'pid' values were removed. For the remaining data, where missing percentages for all features were under 50%, Multi-Imputation (MI) was applied to numerical features with missing values, while missing values in object features were set to '-1'. Additionally, MI was applied to the 'imprace' feature across the entire original dataset of 8,368 records.

After handling outliers and missing values, the two dataset parts are merged, and partner data based on 'pid' is incorporated. New features such as 'sameField', 'sameRegion', 'sameCareer', and 'sameUndergra' are created to assess similarity between individuals and their partners using object features. For numeric features like personality and interests, new features are generated to measure similarity and relationships between individuals and their partners.

The finalized dataset can be referenced at Page 46.

Subsequently, the entire dataset is split into training and test sets for Experiment 1, with 5020 instances in the training set and 3348 instances in the test set. Further splits were made into female and male datasets, with 2510 instances in each gender's training set and 1674 instances in each gender's test set.

4.3 *Algorithms*

Based on the previous work, RF would be chosen as the baseline model. In this study, XGBoost and LightGBM will be used to compare their performance with RF.

4.3.1 *XGBoost*

XGBoost, proposed by Tianqi Chen and Carlos Guestrin in 2016, is a scalable boosting system. It iteratively fits new trees to correct errors made by the current ensemble, gradually improving predictions. Based on traditional gradient boosting, XGBoost introduces advanced features for scalability, speed, and flexibility. These include regularization, out-of-core computation, parallel/distributed computing, an optimized tree learning

algorithm for sparse data, and a weighted quantile sketch procedure. XGBoost's efficiency in computing and its built-in capabilities for handling outliers, missing values, and imbalanced data make it suitable for this study (Chen & Guestrin, 2016).

4.3.2 *LightGBM*

LightGBM, short for Light Gradient Boosting Machine, is a gradient boosting framework known for its faster training speed, higher efficiency, and lower memory usage. It supports parallel, distributed, and GPU learning and handles large-scale data effectively [7]. Compared to traditional Gradient Boosting Decision Tree (GBDT) models like XGBoost, LightGBM achieves similar accuracy but with faster performance, especially on high-dimensional and large datasets. This efficiency is due to Gradient-based One-Side Sampling (GOSS), which excludes data instances with small gradients for faster estimation, and Exclusive Feature Bundling (EFB), which reduces the number of features scanned. LightGBM also employs a histogram-based algorithm for finding the best splits, enhancing decision tree construction and learning speed (Ke et al., 2017; H. Zhang et al., 2017). LightGBM is selected for this study based on its proven ability to achieve nearly the same accuracy as XGBoost but with faster execution times.

4.3.3 *Imbalanced Handling Techniques*

Class weight, SMOTE-ENN, SMOTE-Tomek, random undersampling, and random oversampling will be experimented with in this study to address the imbalanced data.

Class weight is a parameter in the classifiers. In RF, setting 'class_weight' to 'balanced' adjusts the weights to be inversely proportional to class frequencies, according to the values of y. In XGBoost and LightGBM, 'scale_pos_weight' is the parameter that controls the balance of weights for the positive and negative classes.

The other methods—SMOTE-ENN, SMOTE-Tomek, random undersampling, and random oversampling—are resampling techniques. There are two types of resampling: undersampling and oversampling. Undersampling removes instances from the majority class in the training data to balance the majority and minority instances. Oversampling adds similar instances to the minority class to achieve balance. Random oversampling creates and appends additional instances of the minority class to the dataset until the class distribution is balanced. With random undersampling, a subset of the majority class instances is randomly selected and retained while the rest are discarded (He & Garcia, 2009).

[7] https://github.com/microsoft/LightGBM?tab=readme-ov-file

However, resampling may introduce noise and overlap, reducing data quality and affecting model performance. Specifically, due to the randomness in sampling, undersampling may incorrectly remove necessary instances from the majority class, affecting the establishment of classes and construction of discriminant rules for classification, causing noise; it may also retain unnecessary instances, leading to overlap. Conversely, oversampling may create instances that are highly similar to the limited instances in the minority class if the number of instances used in oversampling is extremely limited, leading to overfitting. SMOTE-ENN and SMOTE-Tomek are SMOTE techniques combining both over- and under-sampling methods to handle noise and overlap caused during the resampling process. SMOTE-ENN, short for Synthetic Minority Oversampling Technique and Edited Nearest Neighbor, oversamples the minority class using interpolation and removes redundant instances using the ENN technique to produce balanced data for further model fitting (Muntasir Nishat et al., 2022). SMOTE-Tomek combines SMOTE with Tomek Links to create new instances and identify and remove unnecessary instances to achieve dataset resampling (Batista et al., 2003).

4.3.4 *Hyperparameter Tuning*

Grid search and random search are popular techniques for hyperparameter optimization. Specifically, grid search exhaustively explores a predefined subset of the hyperparameter space, evaluating every possible combination. In contrast, random search randomly samples combinations from the hyperparameter space (Liashchynskyi & Liashchynskyi, 2019). When dealing with a hyperparameter space of the same size, random search is significantly faster than grid search. However, its performance may be slightly lower due to noise, although this does not typically affect out-of-sample performance[8]. Therefore, given time constraints, random search will be used in this study for hyperparameter tuning.

4.4 *Experiments Set-up*

Based on the models and measurements, the experiments will be conducted in sequence to address the research questions.

4.4.1 *Experiment 1*

To assess how well machine learning models (RF, XGBoost, and LightGBM) predict dating matches, RF will serve as the baseline model, known

[8] https://scikit-learn.org/stable/auto_examples/model_selection/plot_randomized_search.html

for its reliability in binary classification tasks from previous studies. XGBoost and LightGBM, both noted for their strong performance, will be compared to determine which model performs best overall. Given the dataset's imbalance, various techniques such as adjusting class weights, using SMOTE-ENN, SMOTE-Tomek, random undersampling, and random oversampling will be applied with these models to find the most effective approach for handling imbalanced data in this prediction task. Subsequently, the baseline model, XGBoost, and LightGBM will be fine-tuned and evaluated on a validation dataset to measure their accuracy in predicting dating matches. Stratified k-fold cross-validation will ensure reliable results during the comparison of imbalanced data handling techniques and hyperparameter tuning. For out-of-sample evaluation, the test set will be divided into 5 folds to provide robust statistical evaluation.

4.4.2 *Experiment 2*

To answer the sub question 2: How well do the models perform when trained separately for males and females, the whole dataset would be split by gender. Since this inquiry focuses on predicting decisions made by female and male candidates (target 'dec' instead of 'match'), models will be separately trained and evaluated on female and male datasets. Imbalanced handling techniques will only be applied to the female dataset, as the male dataset is nearly balanced. Stratified k-fold cross-validation will also be utilized to ensure consistent and reliable evaluation.

4.4.3 *Experiment 3*

The third experiment is conducted to answer the third sub question: Which features are most relevant and important for males and females respectively in predicting dating matches, based on the model with the best performance? It would be based on the second experiment. In this procedure, the best performance model on the testing data in the second experiment would be used with SHAP value to find the important feature to influence the dating candidates' decision.

4.5 *Performance Evaluation*

MCC, BACC and AUC-ROC curves will be utilized as performance evaluation metrics due to the dataset's imbalance, where there are more unmatched instances than matched ones. And MCC would be seen as the main metrics as it is proved to be the robust metrics 3.4. This imbalance makes it crucial to consider both negative and positive instances for this problem. Furthermore, the final models from each experiment will be

compared against the baseline model used in previous studies using test sets to assess their performance. F1-score will also be employed to further compare these models with other classification tasks. Error analysis will involve using confusion matrices to compare models.

4.5.1 *BACC*

Compared to the commonly used Accuracy metric, BACC is more suitable for imbalanced data. According to Equation 1, BACC calculates how many positive instances are correctly classified and how many negative instances are correctly classified, giving them equal weight to compute the final BACC score. In essence, BACC represents the arithmetic mean of TPR and TNR, ensuring that the correct classification of negative instances is equally important as that of positive instances. This characteristic makes BACC suitable for problems where both positive and negative outcomes are equally significant.

$$BalancedAccuracy = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{FP+TN}) \tag{1}$$

4.5.2 *MCC*

MCC, short for Matthews Correlation Coefficient, was introduced by Brian W. Matthews in 1975 and has proven reliable for handling imbalanced data. It measures the correlation between observed and predicted classifications, yielding a high score only when models correctly predict both positive and negative instances. As per Equation 2, MCC considers True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN), providing more comprehensive information than BACC by accounting for FP and FN. MCC ranges from -1 (perfectly wrong prediction) to +1 (perfect prediction), with 0 indicating predictions similar to random guessing (Chicco & Jurman, 2020; Chicco et al., 2021; Tharwat, 2020).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)))}} \tag{2}$$

4.6 *Software*

The interface platform using in this study is Google Colab and the programming language is Python 3.10.12. The packages and libraries used are:

- Pandas (McKinney et al., 2010)
- Numpy (Harris et al., 2020)

- Matplotlib (Hunter, 2007)
- Seaborn (Waskom, 2021)
- Scipy (Virtanen et al., 2020)
- Scikit-learn (Pedregosa et al., 2011)
- XGBoost (Chen & Guestrin, 2016)
- LightGBM (Shi et al., 2024)
- SMOTE (Chawla et al., 2002)
- SHAP (Lundberg & Lee, 2017a)

4.7 *Overview Methodology*

Figure 10 represents the overall workflow in this study.

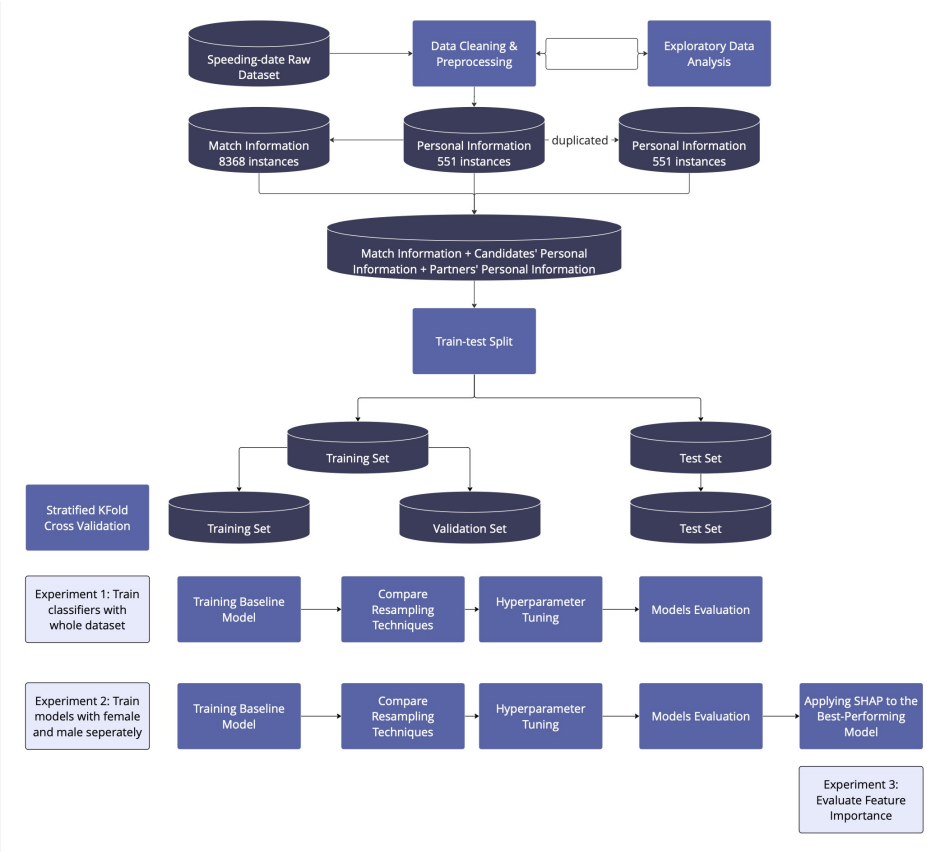


Figure 1: Overview Methodology

5 RESULTS

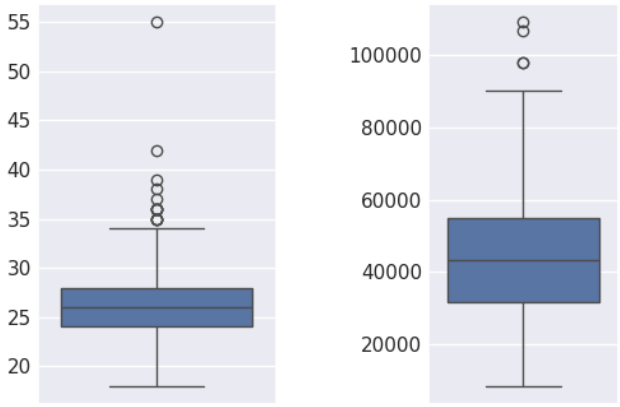
This section reports the result of the exploratory data analysis and experiments. There are three sections of the experiment. First, each model is explored to see their performance on predicting dating matches on the whole dataset, which is experiment 1. Then in experiment 2, models is trained on female and male dataset respectively to see their performance on predicting candidates decision by gender. Finally, in experiment 3, SHAP technique is used with the best performance model in experiment 2 to explore the relevance and important features for female and male dating candidates in making decisions.

5.1 *Exploratory Data Analysis*

In the EDA procedure, features were categorized into several groups based on their data type and meaning. In addition to demographic information, characteristics and interests were also grouped for exploration.

Figure 2a illustrates that most candidates' ages fall within the range of 20-30 years old, with only two candidates beyond 40 years old. This distribution likely reflects the dataset's predominant representation of university students, highlighting dating preferences among younger individuals during that period.

Income is derived from candidates' location and the median household income in that area, rather than their actual income. Figure 2b shows that most candidates' incomes range between 30,000 and 60,000.



(a) Box Plot for Age (b) Box Plot for Income

Figure 2: Box Plots of Age and Income

The personality traits group includes features that describe candidates' self-assessed personality across five dimensions: Attractiveness, Sincerity, Intelligence, Fun, and Ambition. According to Figure 3, all these features range from 1 to 10, indicating no outliers. In comparison to attractiveness, fun, and ambition levels, sincerity and intelligence mostly range between 7 and 10. This suggests that candidates tend to rate themselves higher in sincerity and intelligence.

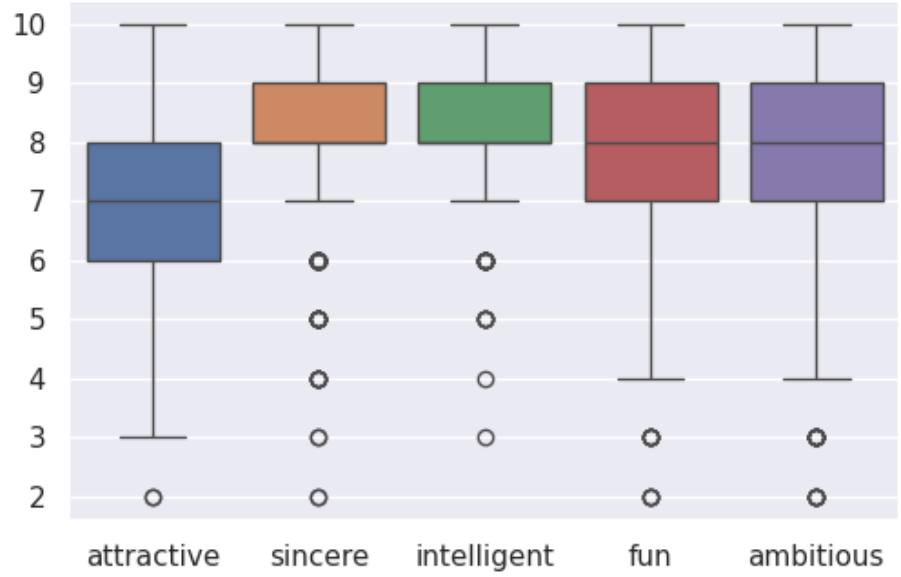


Figure 3: Box plot of Characteristic

The Interest group includes features that describe candidates' interests in various activities like sports, TV shows, exercise, dining, museums, and more. According to Figure 4, reading and gaming are rated beyond the scale of 10, suggesting the need for outlier handling methods. Specifically, some activities are less popular than others, with most activities showing an average interest level between 6 and 8. However, TV sports, gaming, and yoga receive ratings close to or below 4.

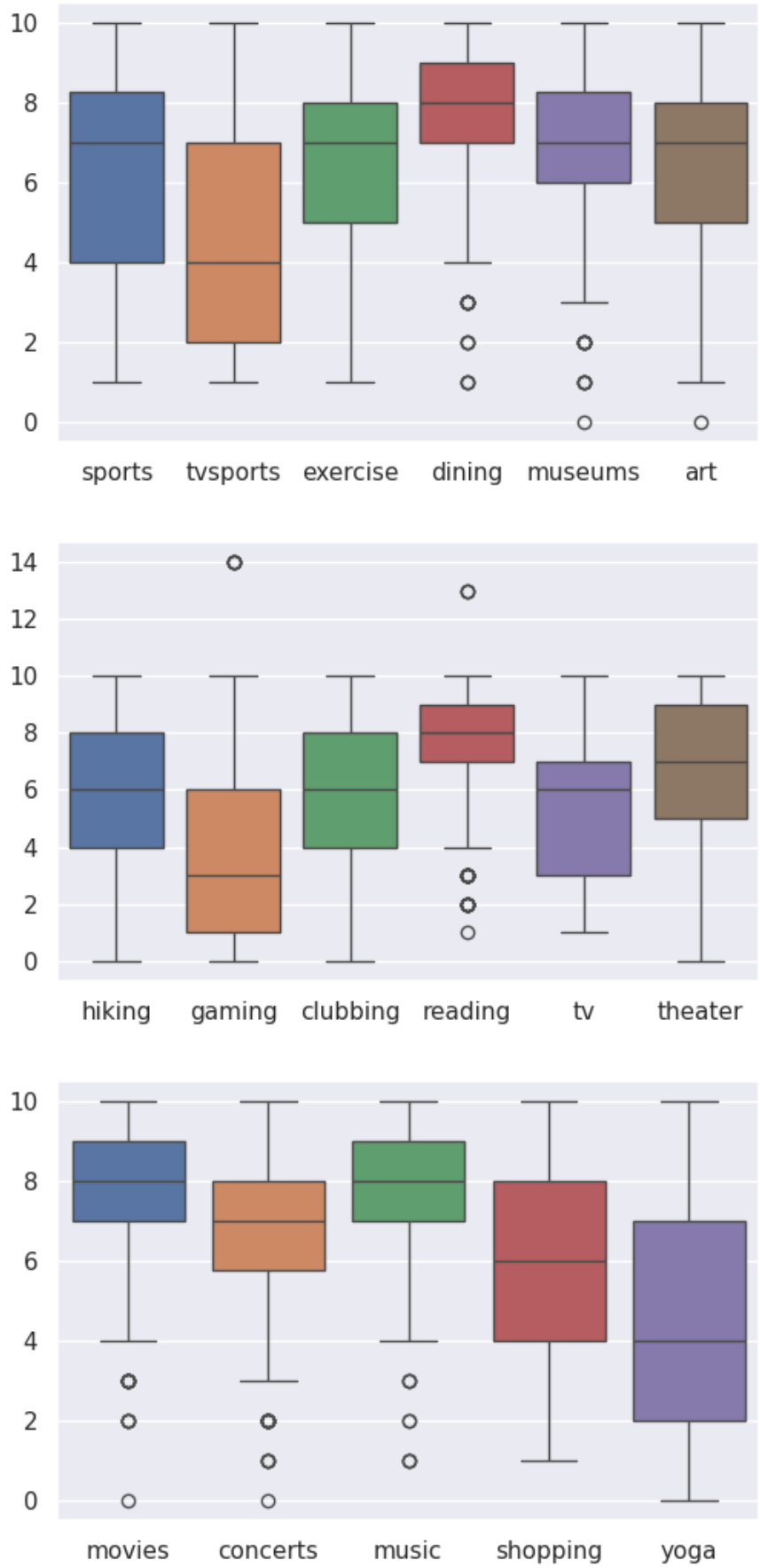


Figure 4: Box plot of Interest

Based on Figure 5 and Figure 6, most candidates either come from or have graduated from the business/economics/finance department. Additionally, they are currently in or planning to pursue careers in business-related fields or academia.

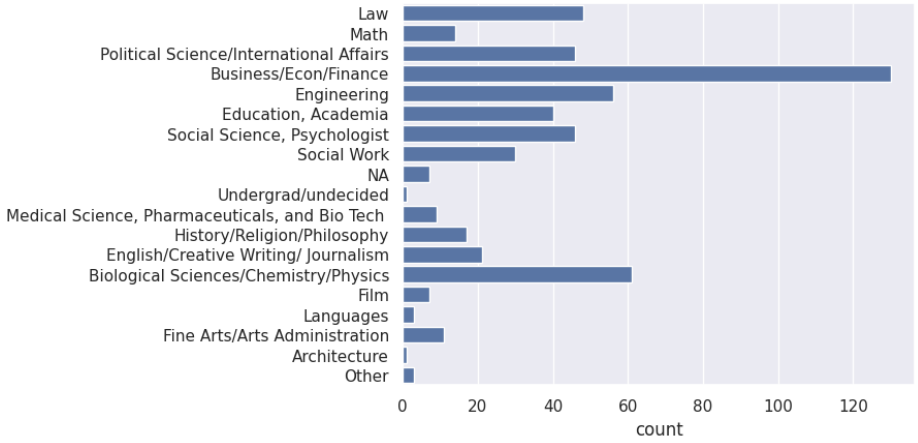


Figure 5: Distribution of Field

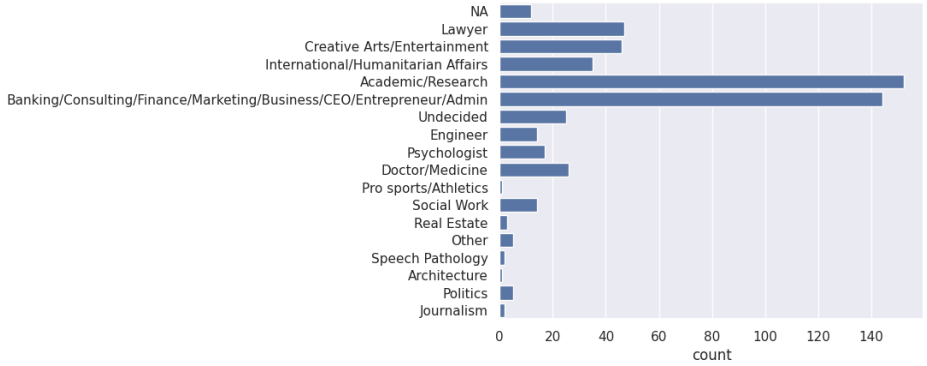


Figure 6: Distribution of Career

5.2 *Results Experiment 1*

The performance of each model in predicting dating matches is presented in Table 1. XGBoost achieves the best performance across all metrics, including F1-score, MCC, and BACC, compared to the baseline model RF and LightGBM.

Table 1: Base Model Comparison

Model	F1-score	MCC	BACC
RF	0.035 ± 0.013	0.090 ± 0.040	0.508 ± 0.004
XGBoost	**0.293 ± 0.031**	**0.309 ± 0.036**	**0.585 ± 0.011**
LightGBM	0.127 ± 0.023	0.168 ± 0.032	0.531 ± 0.007

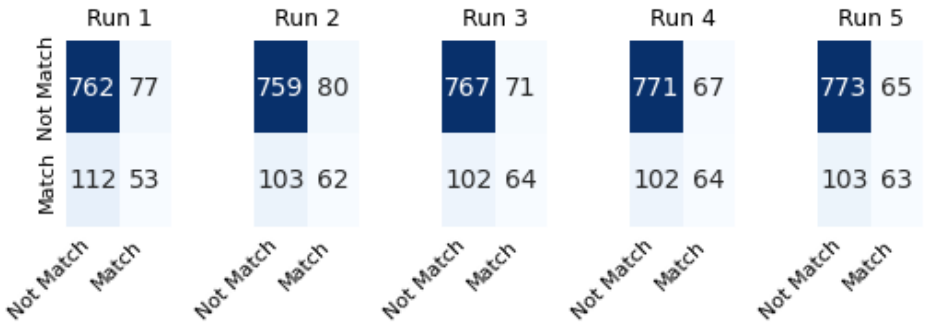
As the dataset is imbalanced, with the majority of the target values being unmatched, multiple imbalance handling methods—including Class weight adjustment, SMOTE-ENN, SMOTE-Tomek, random undersampling, and random oversampling—are applied to explore the best method for each model. According to Table 2, random undersampling performs best with the RF model. For XGBoost, using class weight adjustment yields the best F1-score, MCC, and BACC. When using LightGBM, class weight adjustment achieves the best F1-score, random undersampling achieves the best BACC score, and random oversampling achieves the best MCC score.

Table 2: Model Performance with Imbalanced Handling Methods

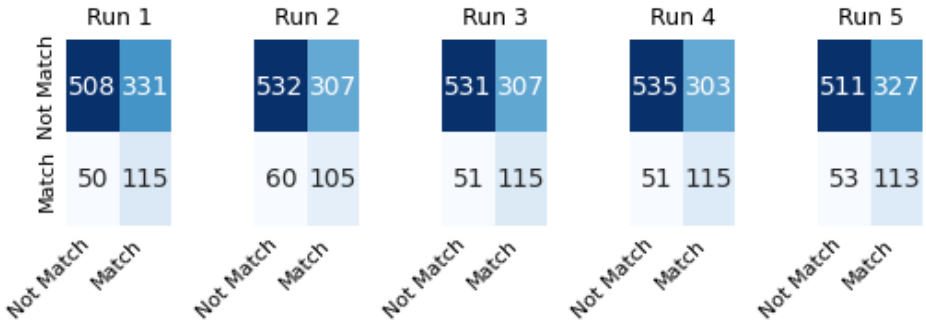
Model	F1-score	MCC	BACC
RF	0.035 ± 0.013	0.090 ± 0.040	0.508 ± 0.004
RF + Class Weight	0.002 ± 0.005	0.000 ± 0.022	0.500 ± 0.001
RF + SMOTE-ENN	0.200 ± 0.051	0.157 ± 0.058	0.548 ± 0.020
RF + SMOTE-Tomek	0.060 ± 0.016	0.083 ± 0.038	0.512 ± 0.005
RF + RandomUnderSampler	**0.374 ± 0.012**	**0.218 ± 0.018**	**0.644 ± 0.012**
RF + RandomOverSampler	0.112 ± 0.032	0.198 ± 0.041	0.529 ± 0.009
XGBoost	0.293 ± 0.031	0.309 ± 0.036	0.585 ± 0.011
XGBoost + Class Weight	**0.423 ± 0.024**	**0.362 ± 0.026**	**0.643 ± 0.012**
XGBoost + SMOTE-ENN	0.263 ± 0.027	0.160 ± 0.033	0.566 ± 0.013
XGBoost + SMOTE-Tomek	0.213 ± 0.056	0.221 ± 0.066	0.556 ± 0.020
XGBoost + RandomUnderSampler	0.365 ± 0.010	0.206 ± 0.016	0.637 ± 0.011
XGBoost + RandomOverSampler	0.401 ± 0.033	0.340 ± 0.035	0.632 ± 0.016
LightGBM	0.127 ± 0.023	0.168 ± 0.032	0.531 ± 0.007
LightGBM + Class Weight	**0.410 ± 0.027**	0.310 ± 0.032	0.642 ± 0.014
LightGBM + SMOTE-ENN	0.255 ± 0.022	0.164 ± 0.020	0.564 ± 0.010
LightGBM + SMOTE-Tomek	0.115 ± 0.031	0.153 ± 0.055	0.527 ± 0.010
LightGBM + RandomUnderSampler	0.380 ± 0.011	0.228 ± 0.017	**0.652 ± 0.011**
LightGBM + RandomOverSampler	0.407 ± 0.023	**0.318 ± 0.026**	0.638 ± 0.012

To determine the most suitable imbalance handling method for LightGBM in this experiment, confusion matrices obtained after fitting LightGBM with each method will be examined, as shown in Figure 7. With random undersampling, the model produced more FN than with class weighting and random oversampling, and tended to incorrectly predict "unmatched" instances as "matched." When applying class weighting with LightGBM, the number of FN decreased, and fewer "unmatched" instances

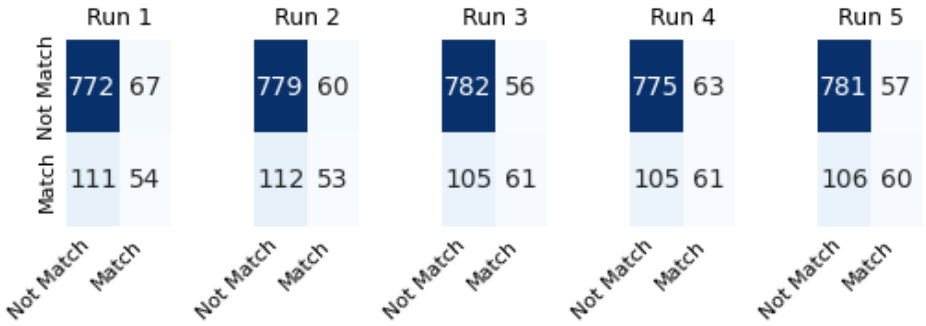
were predicted as "matched." With random oversampling, the model correctly classified an average of 11 more TP than when using class weighting, although it correctly classified 3 fewer TN on average. Therefore, random oversampling will be used with LightGBM for further experiments.



(a) Confusion Matrices for LightGBM ClassWeight



(b) Confusion Matrices for LightGBM RandomUnderSampler



(c) Confusion Matrices for LightGBM RandomOverSampler

Figure 7: Confusion Matrices per Model and Run on LightGBM

With the most suitable imbalance handling methods for each model, hyperparameter tuning is conducted to find the optimal settings and

improve predictive performance. After performing a random search for each model, the optimal hyperparameters are shown in Table 3.

Table 3: Optimal Hyperparameters Set

Model	Optimal Hyperparameter Set
RF	bootstrap = False, criterion = 'gini', max_depth = 17, max_features = 'sqrt', min_samples_leaf = 1, min_samples_split = 6, n_estimators: 109
XGBoost	subsample = 0.90, reg_lambda = 0.001, reg_alpha = 0.701, n_estimators = 350, max_depth = 6, gamma = 0.2, eta = 0.09, colsample_bytree = 0.80
LightGBM	reg_lambda = 0.1, reg_alpha = 0.1, num_leaves = 30, n_estimators = 350, max_depth = 10, learning_rate = 0.15

Each model is retrained using the appropriate imbalance handling methods and optimal hyperparameters. As shown in Table 4, after 5 runs of stratified k-fold cross-validation, each enhanced model achieves higher mean scores in F1-score, MCC, and BACC.

Table 4: Enhanced Model Comparison

Model	F1-score	MCC	BACC
RF	0.374 ± 0.012	0.218 ± 0.018	0.644 ± 0.012
Enhanced RF	0.387 ± 0.017	0.240 ± 0.026	0.659 ± 0.017
XGBoost	0.423 ± 0.024	0.362 ± 0.026	0.643 ± 0.012
Enhanced XGBoost	0.450 ± 0.034	0.425 ± 0.040	0.651 ± 0.015
LightGBM	0.407 ± 0.023	0.318 ± 0.026	0.638 ± 0.012
Enhanced LightGBM	0.437 ± 0.010	0.419 ± 0.010	0.644 ± 0.005

After proving the performance of the enhanced model, the model is fitted using the whole training set and measured with the test set to find their out-of-sample performance. To get a statistical evaluation on out-of-sample performance, the test set is split into 5 folds using a stratified k-fold split. Table 5 shows that LightGBM achieves the best mean on F1-score and MCC and RF achieves the best mean on BACC.

Table 5: Model Performance on Testset

Model	F1-score	MCC	BACC
RF	0.463 ± 0.015	0.354 ± 0.021	**0.732 ± 0.013**
XGBoost	0.543 ± 0.033	0.519 ± 0.045	0.696 ± 0.014
LightGBM	**0.548 ± 0.045**	**0.539 ± 0.049**	0.696 ± 0.022

Based on Figure 8, which shows the ROC curves and AUC scores for each run and each model, it is evident that RF, XGBoost, and LightGBM

all performed better than a random prediction. LightGBM consistently achieved the same or better AUC scores in each run compared to RF and XGBoost.

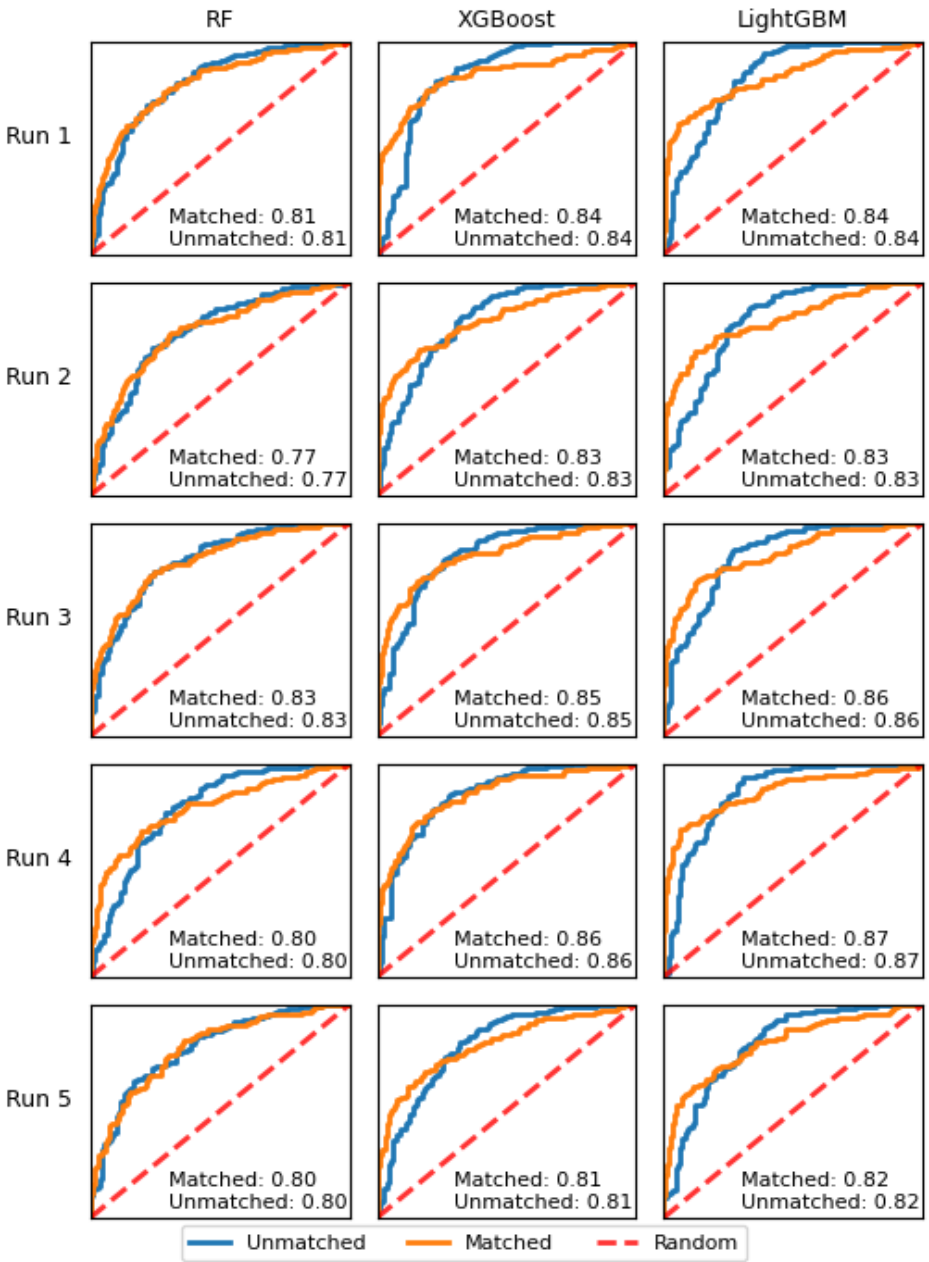


Figure 8: ROC curves and AUC scores per model and run

The Figure 9 presents the results after 5 runs on the test set for each model, indicating the error patterns. It is shown that RF tends to classify

the target value as matched, achieving the highest TN but also tends to wrongly predict unmatched instances as matched, resulting in the highest FN compared to XGBoost and LightGBM, and also the lowest TP. Compared to XGBoost, LightGBM achieves 4 more TP on average while achieving nearly 1 fewer TN on average.

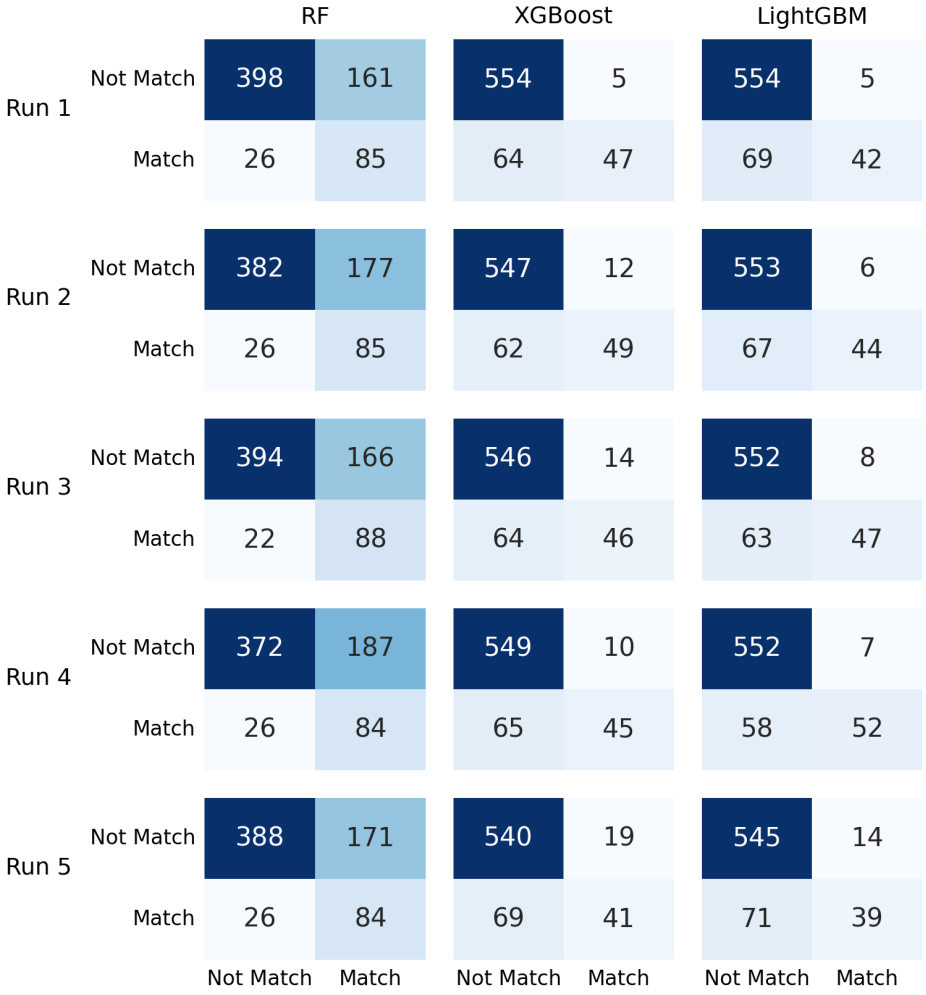


Figure 9: Confusion Matrices per Model and Run

5.3 *Results Experiment 2*

Experiment 2 aims to explore the machine learning models' performance in predicting dating decisions separately for females and males. Table **??** compares the performance of the basic RF, XGBoost, and LightGBM models in predicting the dating decisions made by female and male candidates. In this experiment, the female dataset is imbalanced, though less so than

in Experiment 1, while the male dataset is nearly balanced at a 1:1 ratio. Among all the models, XGBoost achieves the highest average F1-score, and LightGBM performs best on MCC and BACC for the female dataset. For predicting male candidates' decisions, RF achieves the best performance on MCC and BACC, whereas LightGBM scores highest on the F1-score.

Table 6: Base Model Comparison by Gender

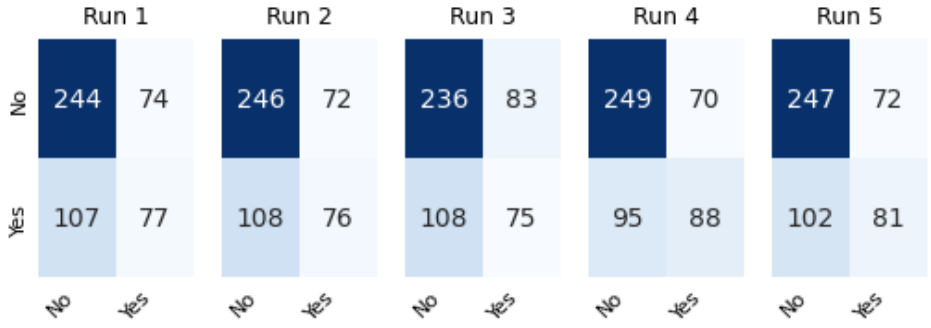
Gender	Model	F1-score	MCC	BACC
Female	RF	0.302 ± 0.029	0.194 ± 0.029	0.565 ± 0.011
	XGBoost	**0.433 ± 0.025**	0.197 ± 0.017	0.589 ± 0.010
	LightGBM	0.428 ± 0.016	**0.210 ± 0.029**	**0.592 ± 0.012**
Male	RF	0.562 ± 0.020	**0.235 ± 0.029**	**0.614 ± 0.014**
	XGBoost	0.571 ± 0.027	0.209 ± 0.041	0.604 ± 0.021
	LightGBM	**0.580 ± 0.011**	0.217 ± 0.029	0.608 ± 0.014

As the female dataset is still imbalanced, the best-imbalanced handling methods will be determined before further experiments. As in experiment 1, class weight, SMOTE-EEN, SMOTE-Tomek, random undersampling, and random oversampling would be applied to all models to be judged on their performance on each model. For RF, random undersampling obtained the best performance with the highest average F1-score, MCC and BACC. For XGBoost, class weight achieves the highest average MCC and BACC, while random oversampling achieves the same average MCC with the same standard deviation so it would also be considered in further comparison. Besides, random undersampling obtains the highest F1-score applying with XGBoost. For LightGBM, class weight gets the highest average MVV and BACC while random undersampling achieves the highest F1-score. Therefore, in further experiments, random undersampling would be applied with RF and the most suitable methods for XGBoost and LightGBM would be further discussed.

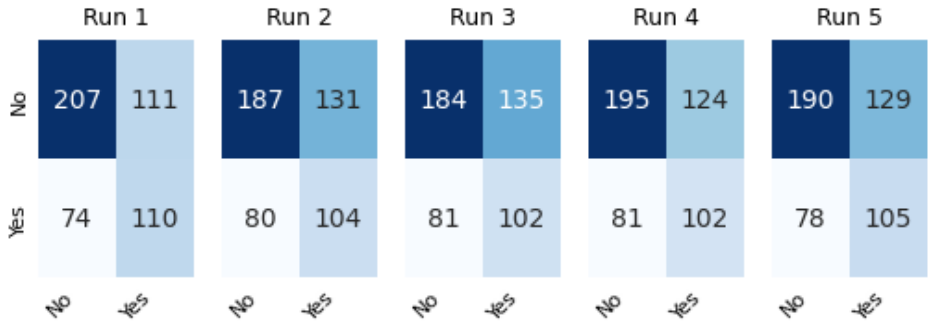
Table 7: Model Performance with Imbalanced Handling Methods on Female Data

Model	F1-score	MCC	BACC
RF	0.302 ± 0.029	0.194 ± 0.029	0.565 ± 0.011
RF + Class Weight	0.242 ± 0.027	0.164 ± 0.049	0.548 ± 0.014
RF + SMOTE-ENN	0.462 ± 0.027	0.155 ± 0.019	0.578 ± 0.011
RF + SMOTE-Tomek	0.343 ± 0.024	0.166 ± 0.033	0.564 ± 0.013
RF + RandomUnderSampler	**0.524 ± 0.009**	**0.203 ± 0.017**	**0.605 ± 0.009**
RF + RandomOverSampler	0.379 ± 0.031	0.197 ± 0.039	0.579 ± 0.016
XGBoost	0.433 ± 0.025	0.197 ± 0.017	0.589 ± 0.010
XGBoost + Class Weight	0.471 ± 0.026	**0.209 ± 0.038**	**0.600 ± 0.019**
XGBoost + SMOTE-ENN	0.460 ± 0.026	0.139 ± 0.030	0.570 ± 0.016
XGBoost + SMOTE-Tomek	0.440 ± 0.032	0.203 ± 0.045	0.592 ± 0.021
XGBoost + RandomUnderSampler	**0.506 ± 0.020**	0.169 ± 0.038	0.587 ± 0.020
XGBoost + RandomOverSampler	0.469 ± 0.019	**0.209 ± 0.038**	0.599 ± 0.017
LightGBM	0.428 ± 0.016	0.210 ± 0.029	0.592 ± 0.012
LightGBM + Class Weight	0.498 ± 0.022	**0.236 ± 0.029**	**0.615 ± 0.014**
LightGBM + SMOTE-ENN	0.462 ± 0.018	0.142 ± 0.023	0.572 ± 0.012
LightGBM + SMOTE-Tomek	0.416 ± 0.032	0.184 ± 0.035	0.582 ± 0.016
LightGBM + RandomUnderSampler	**0.518 ± 0.024**	0.185 ± 0.044	0.596 ± 0.023
LightGBM + RandomOverSampler	0.468 ± 0.028	0.201 ± 0.034	0.597 ± 0.018

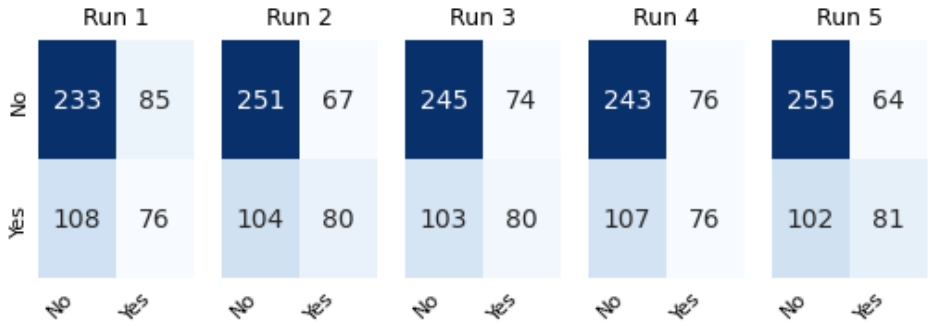
Based on Figure 10, with random undersampling, the model tends to predict the target value as 'Yes', resulting in higher TN and fewer FP. However, it also incorrectly classifies more FN and fewer TP compared to class weighting and random oversampling. Based on the confusion matrices, class weighting and random oversampling yielded quite similar results. For convenience, class weighting will be selected for further experiments with XGBoost, as it does not require additional resampling steps compared to random oversampling but achieves similar results.



(a) Confusion Matrices for XGBoost ClassWeight Female



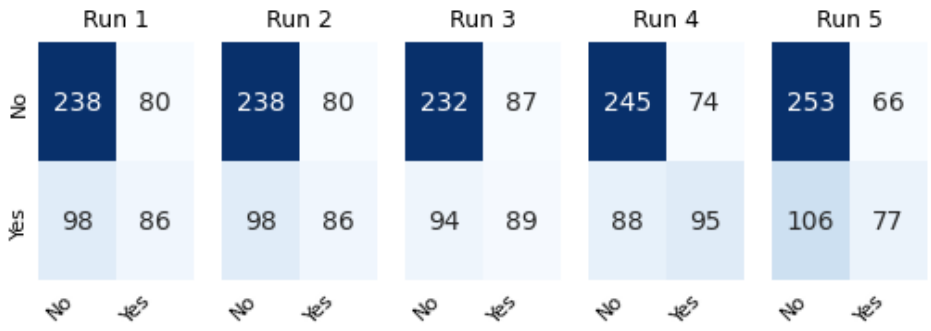
(b) Confusion Matrices for XGBoost RandomUnderSampler Female



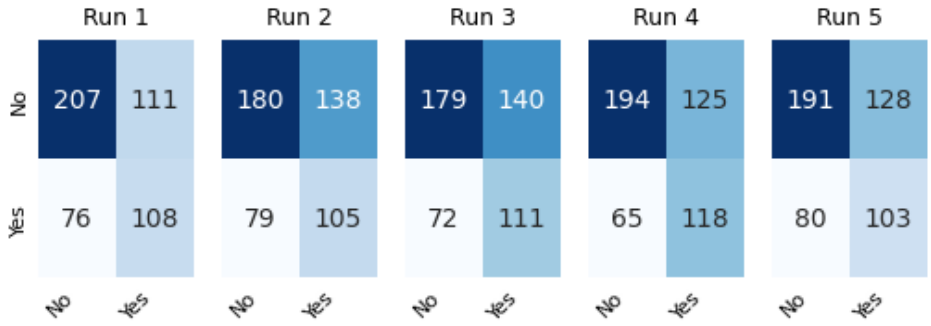
(c) Confusion Matrices for XGBoost RandomOverSampler Female

Figure 10: Confusion Matrices per Model and Run on Female Data XGBoost

Based on Figure 11, with random undersampling, the model tends to predict the target value as 'Yes', resulting in higher TN and fewer FP. However, it also incorrectly classifies more FN and fewer TP compared to class weighting. Therefore, class weighting will be used in further experiments with LightGBM on the female dataset to address the imbalanced problem.



(a) Confusion Matrices for LightGBM with Classweight Female



(b) Confusion Matrices for LightGBM with RandomUnderSampler Female

Figure 11: Confusion Matrices per Model and Run on Female Data LightGBM

After selecting the suitable methods for each model to handle the imbalanced class in the female dataset, random search is conducted to find the optimal hyperparameters for each model for both the female and male datasets, as presented in Table 8.

Table 8: Optimal Hyperparameter Set for Experiment 2

Model	MCC	Optimal Hyperparameter Set
Female	RF	bootstrap = False, criterion = 'gini', max_depth = 19, max_features = 'sqrt', min_samples_leaf = 5, min_samples_split = 7, n_estimators = 188
	XGBoost	subsample = 0.90, reg_lambda = 0.401, reg_alpha = 0.301, n_estimators = 450, max_depth = 4, gamma = 0.4, eta = 0.13, colsample_bytree = 0.6
	LightGBM	reg_lambda = 100, reg_alpha = 1, num_leaves = 45, n_estimators = 500, max_depth = 6, learning_rate = 0.15
Male	RF	bootstrap = False, criterion = 'entropy', max_depth = 17, max_features = 'sqrt', min_samples_leaf = 1, min_samples_split = 15, n_estimators = 87
	XGBoost	subsample = 0.80, reg_lambda = 0.401, reg_alpha = 0.401, n_estimators = 450, max_depth = 7, gamma = 0.40, eta = 0.01, colsample_bytree = 0.50
	LightGBM	reg_lambda = 50, reg_alpha = 2, num_leaves = 30, n_estimators = 450, max_depth = 9, learning_rate = 0.2

Then the model is trained again with the optimal hyperparameters, and the results after stratified 5-fold cross-validation are shown in Table 9 and Table 10. These two tables show that after using the optimal hyperparameters, all the enhanced models achieve higher average F1-scores, MCC, and BACC compared to the basic models.

Table 9: Enhanced Model Comparison on Female Data

Model	F1-score	MCC	BACC
RF	0.524 ± 0.009	0.203 ± 0.017	0.605 ± 0.009
Enhanced RF	0.535 ± 0.019	0.220 ± 0.038	0.613 ± 0.019
XGBoost	0.471 ± 0.026	0.209 ± 0.038	0.600 ± 0.019
Enhanced XGBoost	0.510 ± 0.038	0.258 ± 0.047	0.625 ± 0.025
LightGBM	0.498 ± 0.022	0.236 ± 0.029	0.615 ± 0.014
Enhanced LightGBM	0.508 ± 0.028	0.245 ± 0.039	0.620 ± 0.020

Table 10: Enhanced Model Comparison on Male Data

Model	F1-score	MCC	BACC
RF	0.562 ± 0.020	0.235 ± 0.029	0.614 ± 0.014
Enhanced RF	0.576 ± 0.010	0.252 ± 0.028	0.623 ± 0.013
XGBoost	0.571 ± 0.027	0.209 ± 0.041	0.604 ± 0.021
Enhanced XGBoost	0.589 ± 0.012	0.257 ± 0.015	0.627 ± 0.007
LightGBM	0.580 ± 0.011	0.217 ± 0.029	0.608 ± 0.014
Enhanced LightGBM	0.587 ± 0.010	0.236 ± 0.018	0.617 ± 0.009

Then the female test set and male test set are split into 5 folds to evaluate the models' generalization. Table 11 shows the results evaluating the performance of the models with optimal hyperparameters, trained on the female and male training sets respectively, and tested on the corresponding test sets for each fold, presented as mean ± standard deviation. For the female dataset, LightGBM performs well in out-of-sample evaluation, achieving the highest F1-score, MCC, and BACC. For the male dataset, RF achieves the best F1-score and BACC. Although XGBoost gets a slightly higher average MCC than RF, it has a larger standard deviation.

Table 11: Model Performance by Gender on Test Set

Gender	Model	F1-score	MCC	BACC
Female	RF	0.578 ± 0.025	0.294 ± 0.039	0.652 ± 0.020
	XGBoost	0.541 ± 0.035	0.284 ± 0.052	0.641 ± 0.026
	LightGBM	**0.587 ± 0.018**	**0.318 ± 0.033**	**0.663 ± 0.017**
Male	RF	**0.655 ± 0.014**	0.334 ± 0.029	**0.667 ± 0.015**
	XGBoost	0.631 ± 0.019	**0.336 ± 0.032**	0.666 ± 0.015
	LightGBM	0.633 ± 0.019	0.313 ± 0.037	0.656 ± 0.018

Based on Figure 12, all the models achieve better performance than a random prediction in each run on predicting the female decision. XGBoost achieves equal or slightly higher AUC scores for each run than RF except for Run 5. LightGBM performs better in Run 1 and especially in Run 3 than RF and XGBoost. LightGBM achieves higher average AUC scores of 0.714, compared to RF's 0.706 and XGBoost's 0.710.

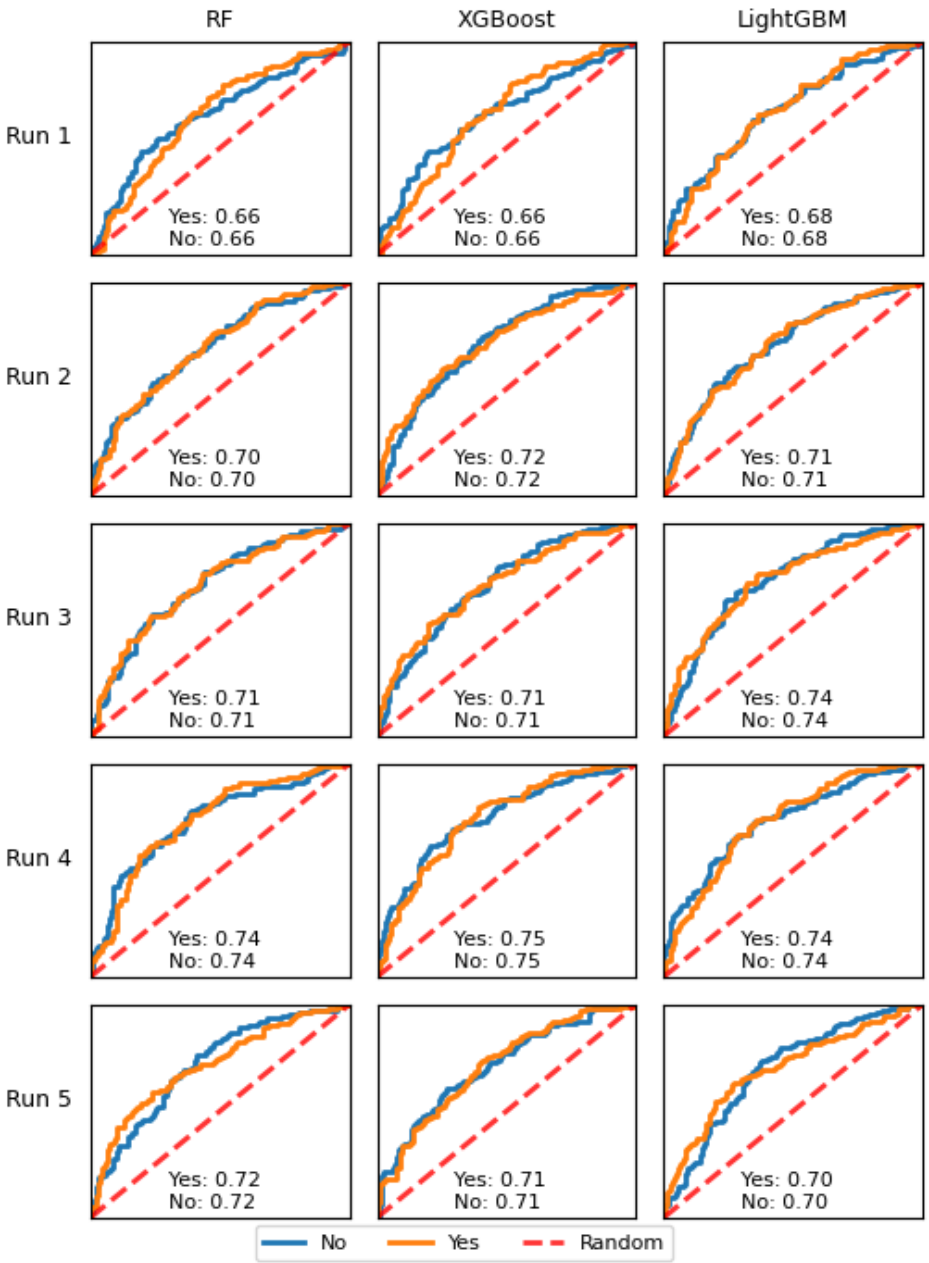


Figure 12: ROC curves and AUC scores per model and run on Female Data

The Figure 13 presents the error patterns on the female test set. It indicates that both RF and LightGBM tend to predict 'Yes' more frequently for the target value. However, LightGBM produces more TP and fewer FN in each run than RF, with an average of nearly 8 more TP, though it produces fewer TN except for Run 1 and Run 2, and on average 2 fewer TN. Compared to LightGBM, XGBoost produces fewer FN and more TP,

with an average of 13 more TP, but it also produces more FP and fewer TN, with an average of 13 fewer TN.

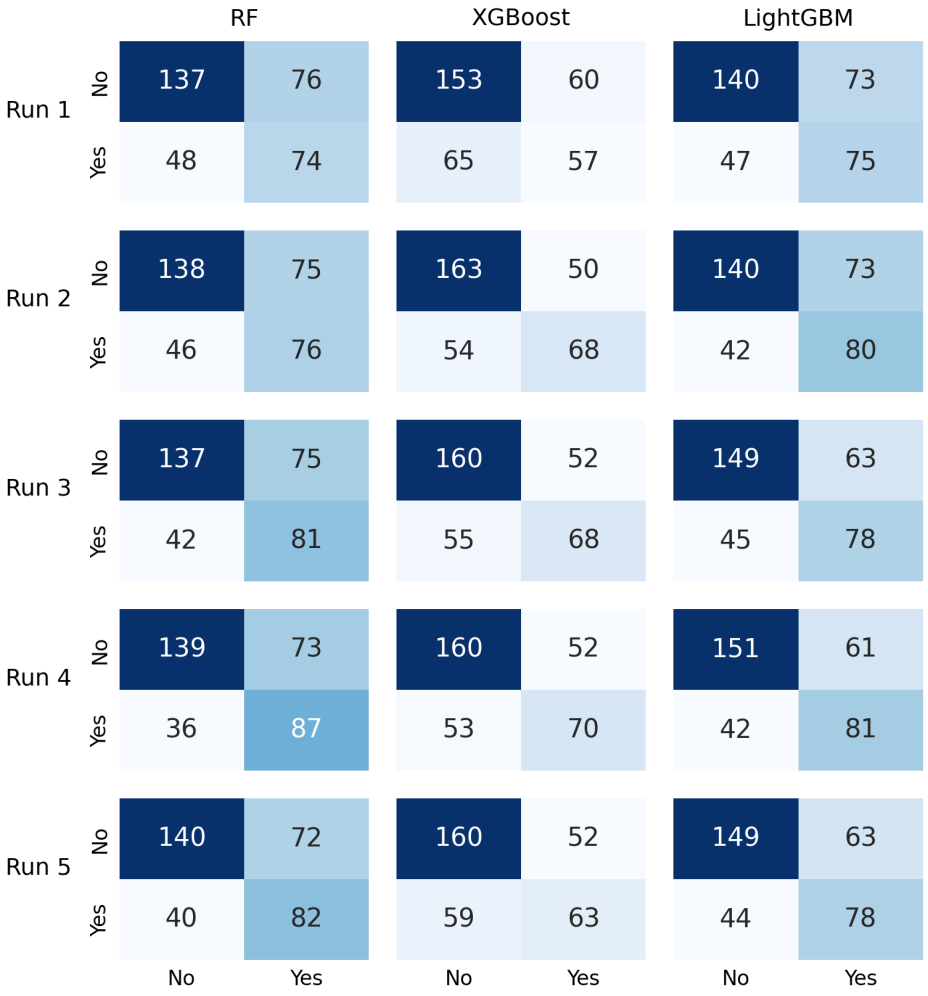


Figure 13: Confusion Matrices per Model and Run on Female Data

The Figure 14 indicates that in predicting the male candidates' decisions, each model achieves better performance than random prediction in each run. Specifically, LightGBM obtains equal or lower AUC scores for each run compared to RF and XGBoost, averaging a score of 0.708. XGBoost achieves a higher average AUC score than RF, with scores of 0.726 and 0.724, respectively. XGBoost's results are also more robust, with a lower standard deviation compared to RF, at 0.019 and 0.023, respectively.

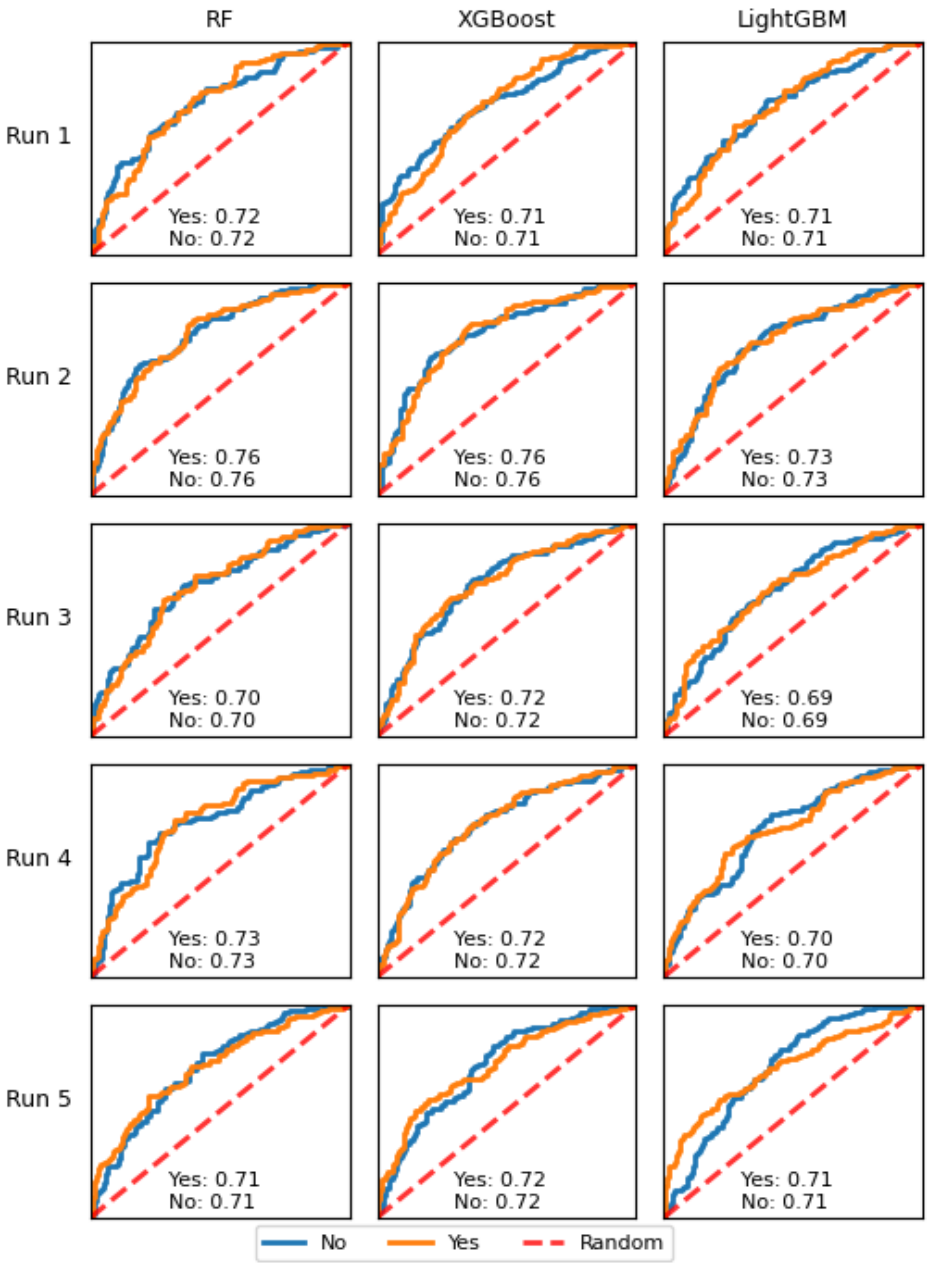


Figure 14: ROC curves and AUC scores per model and run on Male Data

The Figure 15 demonstrates the error patterns when predicting the male candidates' decisions on the test set. It reveals that compared to XGBoost and LightGBM, RF tends to predict the target value as 'No', producing more TN and fewer FP. On average, RF has 10 more TN than XGBoost, but also more FN and slightly fewer TP except for Run 2, averaging 11 fewer TP than XGBoost. Compared with LightGBM, XGBoost produces more TP

and fewer FN, averaging 8 more TP, but also equal or more FP and fewer TN, averaging 4 fewer TN.

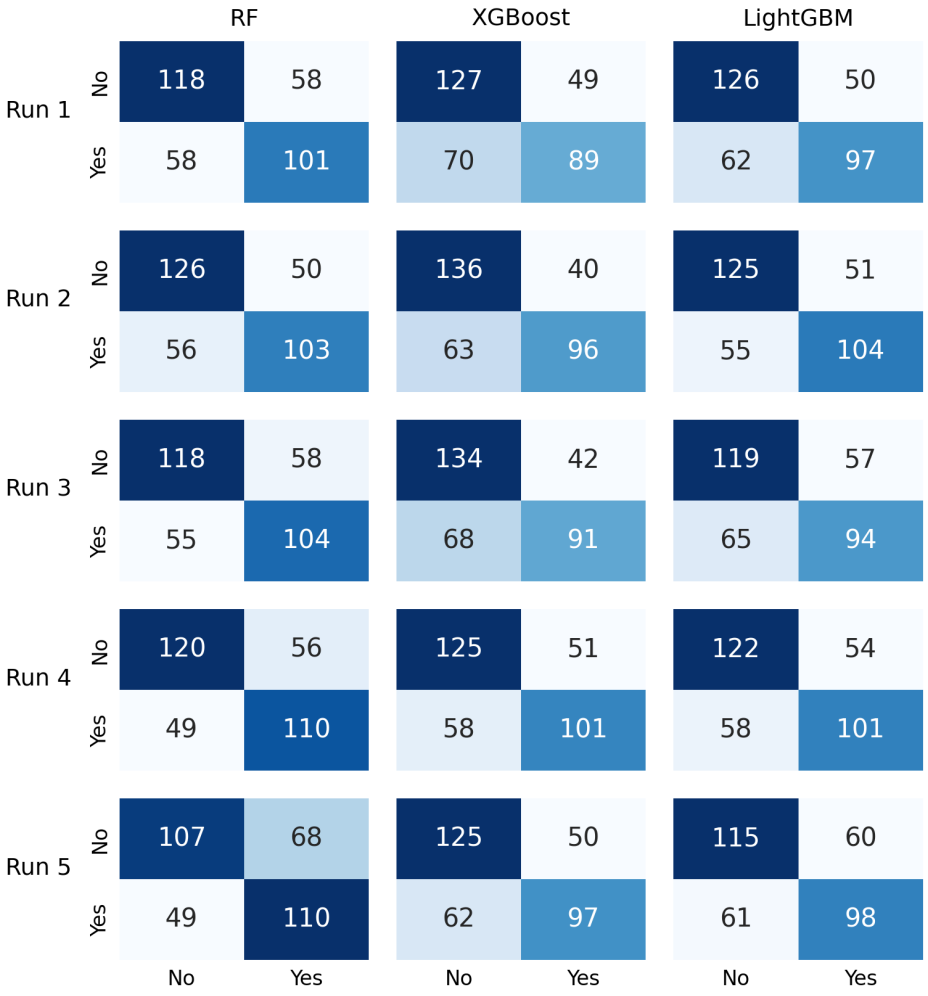


Figure 15: Confusion Matrices per Model and Run on Male Data

5.4 *Results Experiment 3*

In this section, the model with the highest MCC score on the test set, as seen from the results in Experiment 2, will be used with SHAP to display the top 10 features influencing decision-making for female and male candidates, respectively.

5.4.1 *Key Features for Female's Dating Decision*

Figure 16 displays the top 10 features with relatively high SHAP values from the model that achieved the best MCC score in predicting women's dating decisions, specifically LightGBM. The income of their potential partner, their own income, and their potential partner's fun level are the top 3 features influencing female dating decisions. Other significant features include their own attractiveness level, racial similarity with their potential partner and its importance to them, differences in interest in Yoga, their own age, income disparity with their potential partner, their potential partner's age, and their potential partner's attractiveness level.

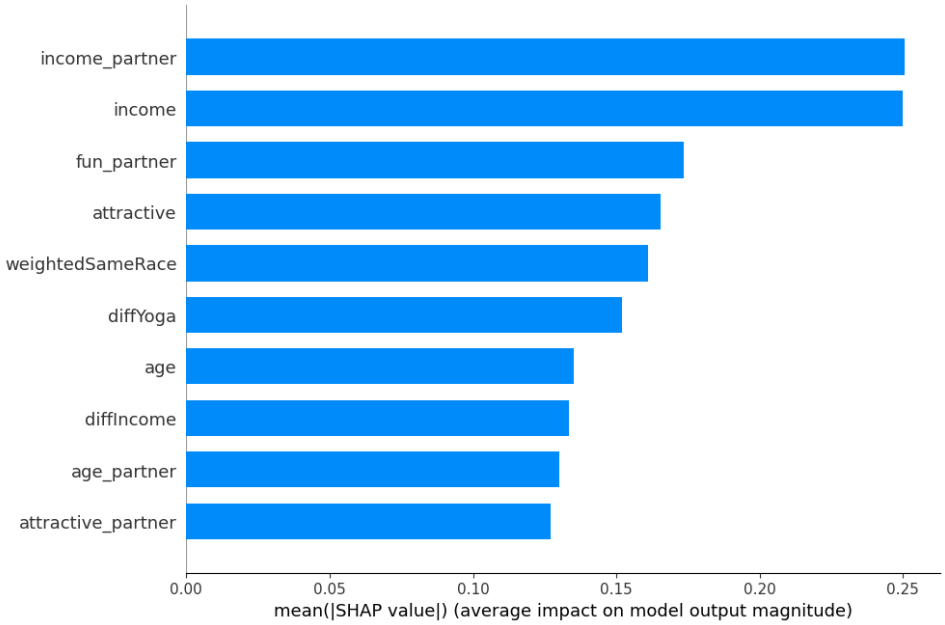


Figure 16: SHAP Summary Plot in Predicting Female's Decision

5.4.2 *Key Features for Male's Dating Decision*

Figure 17 shows the top 10 features with relatively high SHAP values obtained from XGBoost in predicting men's dating choices. Specifically, the attractiveness level of their potential partner, their own income, and the age of their potential partner are the top 3 features influencing male dating decisions. Additionally, factors such as the income of their potential partner, their own ambition level, whether they and their potential partner attended the same undergraduate school or studied the same field, their age difference, and their interests in reading and movies also influence male dating decisions.

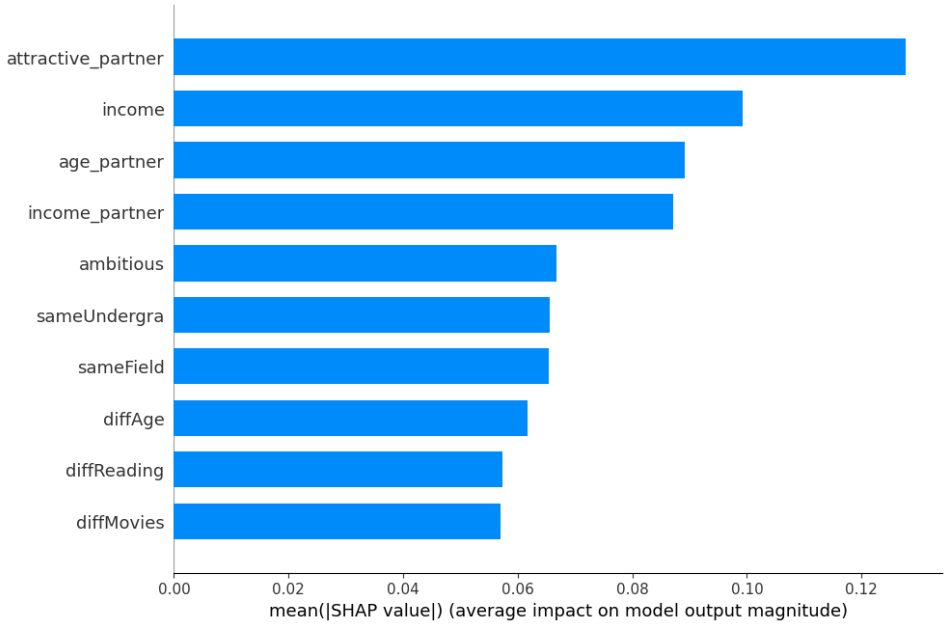


Figure 17: SHAP Summary Plot in Predicting Male's Decision

6 DISCUSSION

In this section, the results of the experiment for each research question, the limitations of this study, its relevance to related work and societal impact, and possible improvements for future work will be discussed.

6.1 *Results Discussion*

6.1.1 *Sub-RQ1: What is the performance of machine learning models in predicting dating matches?*

In predicting overall dating matches on the unseen test set, the enhanced LightGBM with random oversampling achieved an MCC score of 0.539 and an F1-score of 0.548, with the same or better AUC score in each run. Based on the confusion matrices, LightGBM and XGBoost provide a better balance in classification performance compared to RF. LightGBM achieves 4 more TP on average and just 1 fewer TN than XGBoost. Overall, LightGBM exceeds the performance of XGBoost and the baseline model, which is RF.

During the experiment, multiple imbalance handling methods were applied to these three models. Contrary to previous work indicating that SMOTE-related techniques produce better performance, random undersampling, random oversampling, and class weight achieved higher scores with these models, respectively. This may be due to the complexity of the

raw training set, making it difficult to construct clear class discriminant rules. Consequently, SMOTE techniques struggled to create or remove distinct instances to prevent overlap and noise, thereby increasing the training set's complexity. This led to a decrease in the performance of models combined with SMOTE techniques compared to models with their inherent regularization features (Batista et al., 2003, 2004; Fernández et al., 2018; Sasada et al., 2020).

6.1.2 *Sub-RQ2: How well do the models perform when trained separately for males and females?*

When predicting female dating candidates' decisions on the unseen test set, LightGBM with class weight achieved an average F1-score of 0.587, an MCC score of 0.318, and a BACC score of 0.663. It also achieved a higher average AUC score of 0.714, outperforming the baseline model, RF, and XGBoost. Overall, LightGBM performed better.

When predicting male candidates' decisions, the baseline model, RF, achieved a higher average F1-score and BACC score, which were 0.655 and 0.667, respectively. XGBoost obtained the highest average MCC score but with a larger standard deviation than RF. Additionally, XGBoost produced a higher average AUC score of 0.726 compared to RF.

To select a model for the next experiment to obtain feature importance, XGBoost was chosen for predicting male candidates' decisions due to its higher MCC and more balanced classification compared to RF. For predicting female candidates' decisions, LightGBM was selected based on its better performance.

6.1.3 *Sub-RQ3: Which features are most relevant and important for males and females respectively in predicting dating matches, based on the model with the best performance?*

Aligning with the previous work in section 3.1, the income of a potential partner is the most important feature for females in dating selection, while the attractiveness level of a potential partner is the most important for male candidates. Additionally, for female candidates, their own income holds vital importance, similar to their potential partner's income. Among personality features, the fun level of their potential partner is the most important for females. For male candidates, the age of their potential partner is also among the top three important features, aligning with their focus on attractiveness. Furthermore, compared to female candidates, males care more about the study background of their potential partner, such as whether they attended the same undergraduate school or studied in the same field.

6.1.4 *RQ: To what extent can machine learning models predict the suitable match among various dating candidates?*

Based on the findings of this study, it is proven that machine learning models are suitable for predicting dating matches and the decisions of female and male dating candidates. Specifically, according to the F1-score, MCC, BACC, and AUC scores, and the prediction balance, LightGBM with random oversampling performs the best in predicting dating matches. LightGBM with class weight performs the best in predicting female candidates' dating decisions, while XGBoost performs the best in predicting male candidates' dating decisions. The SHAP technique is utilized to improve the model's interpretability in dating candidates' decision-making and to reveal the key features influencing their dating decisions. The income and fun level of a potential partner, along with their own income, are the top three important features for female candidates. For male candidates, the attractiveness level, age of their potential partner, and their own income are the top three important features in dating selection.

6.2 *Limitations*

This study has several acknowledged limitations. Firstly, the dataset used was collected between 2002-2004, which may not accurately represent current dating preferences. Additionally, the speed dating experiment primarily involved participants around 25 years old, focusing predominantly on younger generations and potentially limiting its applicability to a broader population. Moreover, the study solely concentrated on features available in the original speed dating experiment, which may not align with the features commonly used in modern dating apps.

Furthermore, the underlying theory behind dating candidates' decision-making remains unclear, despite identifying key factors that influence their choices. For females, these include their own income, their potential partner's income, and their potential partner's fun level. For males, these factors include their own income, the age of their potential partner, and their potential partner's attractiveness level.

6.3 *Relevance*

As mentioned in Section 3, this study applies RF, XGBoost, and LightGBM with various techniques for handling imbalanced data to predict dating matches and candidates' decisions. It focuses on datasets with limited portfolio features, which presents a novel approach compared to previous research. Unlike earlier studies (Fernández et al., 2018), SMOTE-ENN and

SMOTE-Tomek did not perform as expected in predicting dating matches and female candidates' decisions when compared to untreated data.

Furthermore, the study identifies key features critical for candidates' decision-making, reinforcing existing findings while uncovering new insights. For instance, it underscores the significance of candidates' own income for females and the educational background of potential partners for males, aspects that have received limited attention in prior research.

Moreover, enhancing the accuracy of predicting dating matches and decisions through this study could improve the efficacy of matchmaking in both online and offline dating settings. This could facilitate smoother interactions between candidates based on shared interests and similarities, potentially leading to stronger relationships with increased satisfaction. By reducing the time and effort individuals invest in finding compatible partners, the study aims to contribute to overall happiness and well-being in society.

Additionally, beyond its applications in dating, the methodologies employed in this study could be adapted to address other binary classification and matching challenges, such as labor market placements and college admissions. Furthermore, the model's ability to predict candidates' dating decisions opens doors for behavioral studies focused on understanding the importance of features in human decision-making processes.

In conclusion, this study not only advances predictive modeling in dating but also holds promise for improving decision-making processes across various societal contexts. Its ultimate goal is to enhance personal well-being and promote societal harmony through more effective matchmaking and decision support systems.

6.4 *Future Work*

Based on the findings and limitations of this study, there are several areas for potential improvement and subjects for further research.

Firstly, future studies on dating matches could benefit from using datasets collected from a broader population and dating apps, which would better reflect current dating selection behaviors.

Moreover, while the MCC scores achieved in this study surpass the baseline model and are comparable to previous research (Zang et al., 2017a), they do not exceed 0.6 in predicting both overall dating matches and candidates' decision-making. This suggests there is still room to enhance predictive performance.

Finally, this study identifies key features influencing decision-making, such as candidates' own income, which has not been thoroughly explored

in prior research. Sociological and behavioral studies could further investigate these factors in dating decision-making processes.

7 CONCLUSION

This study explores how well machine learning models (Random Forest, XGBoost, and LightGBM) predict suitable matches among dating candidates and understand decision-making among female and male candidates. Unlike previous studies, it uses these models with various techniques for handling imbalanced data, evaluating them based on F1-score, MCC, and BACC metrics. LightGBM with random oversampling emerges as the top performer in predicting overall dating matches, showing superior metrics like F1-score, MCC, BACC, and AUC, thereby ensuring better classification balance. For predicting decisions of female candidates, LightGBM with class weight excels, while XGBoost outperforms in MCC scores and AUC values for male candidates. SHAP analysis reveals critical factors influencing decisions, such as income and fun levels for females, and attractiveness and age for males. The study acknowledges limitations due to the dataset's age and participant demographics, which may restrict its ability to fully capture current dating preferences across a broader population. Nevertheless, it emphasizes its scientific and societal relevance by aiming to enhance relationship satisfaction and personal happiness. It suggests scaling its methodologies to solve broader binary classification and matching challenges. Future research could improve these predictive models by integrating more extensive demographic datasets and exploring deeper sociological factors.

REFERENCES

Aderoju, S. A., & Jolayemi, E. T. (2019). Issues of class imbalance in classification of binary data: A review.

Alomrani, M. A., Moravej, R., & Khalil, E. B. (2021). Deep policies for online bipartite matching: A reinforcement learning approach. *arXiv preprint arXiv:2109.10380*.

Batista, G. E., Bazzan, A. L., Monard, M. C., et al. (2003). Balancing training data for automated annotation of keywords: A case study. *Wob*, *3*, 10–8.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29.

Bifarin, O. O. (2023). Interpretable machine learning with tree-based shapley additive explanations: Application to metabolomics datasets for binary classification. *Plos one*, *18*(5), e0284315.

Buss, D. M., & Schmitt, D. P. (2019). Mate preferences and their behavioral manifestations. *Annual review of psychology*, *70*, 77–110.

Canbek, G., Taskaya Temizel, T., & Sagiroglu, S. (2022). Ptopi: A comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. *SN Computer Science*, *4*(1), 13.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*, 1–13.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *Ieee Access*, *9*, 78368–78381.

Cruces, J. M. S., Hawrylak, M. F., & Delegido, A. B. (2015). Interpersonal variability of the experience of falling in love. *International Journal of Psychology and Psychological Therapy*, *15*(1), 87–100.

Danquah, R. A. (2020). Handling imbalanced data: A case study for binary class problems. *arXiv preprint arXiv:2010.04326*.

DEMETER, E., & PAIUSAN, A.-M. (2018). Intimate relationship experience effects on individual an psychological tendencies within an intimate

relationship: Preliminary investigation. *Agora Psycho-Pragmatica*, *12*(1), 78–91.

Dietrich, D. (2016). The psychology of romantic relationships.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863–905.

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.

Flach, P. (2019). Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 9808–9814.

Furnham, A., & Tsoi, T. (2012). Personality, gender, and background predictors of partner preferences. *North American Journal of Psychology*, *14*(3).

Gale, D., & Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, *69*(1), 9–15.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284.

Howe, F. (2002). The value of intimate relationships and the challenge of conflict. *Journal of Invitational Theory and Practice*, *8*, 15–26.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(03), 90–95.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Keser, S. B., & Keskin, K. (2023). A gradient boosting-based mortality prediction model for covid-19 patients. *Neural Computing and Applications*, *35*(33), 23997–24013.

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: A big comparison for nas. *arXiv preprint arXiv:1912.06059*.

Lundberg, S. M., & Lee, S.-I. (2017a). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Lundberg, S. M., & Lee, S.-I. (2017b). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

McKinney, W., et al. (2010). Data structures for statistical computing in python. *SciPy*, *445*(1), 51–56.

Michèle, B., & Francesconi, M. (2006). Can anyone be" the'one? evidence on mate selection from speed dating. *Forschungsinstitut zur Zukunft der Arbeit/Institute for the Study of Labor: IZA Discussion Paper Series; 2377*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mühlbauer, S., & Weber, E. (2022). *Machine learning for labour market matching* (tech. rep.). IAB-Discussion Paper.

Muniruzzaman, M. (2017). Transformation of intimacy and its impact in developing countries. *Life sciences, society and policy*, *13*, 1–19.

Muntasir Nishat, M., Faisal, F., Jahan Ratul, I., Al-Monsur, A., Ar-Rafi, A. M., Nasrullah, S. M., Reza, M. T., & Khan, M. R. H. (2022). A comprehensive investigation of the performances of different machine learning classifiers with smote-enn oversampling technique and hyperparameter optimization for imbalanced heart failure dataset. *Scientific Programming*, *2022*, 1–17.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rao, S., Mehta, S., Kulkarni, S., Dalvi, H., Katre, N., & Narvekar, M. (2022). A study of lime and shap model explainers for autonomous disease predictions. *2022 IEEE Bombay Section Signature Conference (IBSSC)*, 1–6.

Ravindranath, S. S., Feng, Z., Li, S., Ma, J., Kominers, S. D., & Parkes, D. C. (2023). Deep learning for two-sided matching.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Sasada, T., Liu, Z., Baba, T., Hatano, K., & Kimura, Y. (2020). A resampling method for imbalanced datasets considering noise and overlap. *Procedia Computer Science*, *176*, 420–429.

Schindler, I., Fagundes, C. P., & Murdock, K. W. (2010). Predictors of romantic relationship formation: Attachment style, prior relationships, and dating goals. *Personal Relationships*, *17*(1), 97–105.

Shi, Y., Ke, G., Soukhavong, D., Lamb, J., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., Titov, N., & Cortes, D. (2024). *Lightgbm: Light gradient boosting machine* [R package version 4.3.0.99]. https://github.com/Microsoft/LightGBM

Singh, P., & Kumar, R. (2023). Assessing imbalanced datasets in binary classifiers. In *Soft computing for problem solving: Proceedings of the socpros 2022* (pp. 291–303). Springer.

Siraj-Ud-Doulah, M., & Islam, M. N. (2023). Performance evaluation of machine learning algorithm in various datasets.

Tharwat, A. (2020). Classification assessment methods. *Applied computing and informatics*, *17*(1), 168–192.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Xia, P., Liu, B., Sun, Y., & Chen, C. (2015). Reciprocal recommendation system for online dating. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 234–241.

Zang, X., Yamasaki, T., Aizawa, K., Nakamoto, T., Kuwabara, E., Egami, S., & Fuchida, Y. (2017a). How competitive are you: Analysis of people's attractiveness in an online dating system. *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 973–978.

Zang, X., Yamasaki, T., Aizawa, K., Nakamoto, T., Kuwabara, E., Egami, S., & Fuchida, Y. (2017b). You will succeed or not? matching prediction in a marriage consulting service. *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 109–116.

Zhang, H., Si, S., & Hsieh, C.-J. (2017). Gpu-acceleration for large-scale tree boosting. *arXiv preprint arXiv:1706.08359*.

Zhang, L., Wang, X., & Yamasaki, T. (2019). Deep feature interaction embedding for pair matching prediction. In *Proceedings of the acm multimedia asia* (pp. 1–6).

APPENDIX A

Table 12: Description of features

Attributes	Description
age	object's age
age_o	partner's age
diffAge	difference between object and partner's age
weightedsameRace	if object and partner in the same race (1,0)
sameRegion	if object and partner come from the same region (1,0)
sameField	if object and partner in the same study field (1,0)
diffSports	difference between object and partner's interests in sports
diffTvsports	difference between object and partner's interests in tvsports
diffExercise	difference between object and partner's interests in exercise
diffDining	difference between object and partner's interests in dining
diffMuseums	difference between object and partner's interests in museums
diffArt	difference between object and partner's interests in art
diffHiking	difference between object and partner's interests in hiking
diffGaming	difference between object and partner's interests in gaming
diffClubbing	difference between object and partner's interests in clubbing
diffReading	difference between object and partner's interests in reading
diffTv	difference between object and partner's interests in Tv
diffTheater	difference between object and partner's interests in theater
diffMovies	difference between object and partner's interests in movies
diffConcerts	difference between object and partner's interests in concerts
diffMusic	difference between object and partner's interests in music
diffShopping	difference between object and partner's interests in shopping
diffYoga	difference between object and partner's interests in yoga
diffFun	difference between object and partner's fun level
sameUndergra	if object and partner come from the same undergraduate school (1,0)
sameCareer	if object and partner in the same career field (1,0)
career	object's career field
career_o	partner's careerfield
income	object's career field
income_o	partner's careerfield
diffIncome	difference between object and partner's income
diffAmb	difference between object and partner's ambition level
diffIntel	difference between object and partner's intelligence level
diffAttr	difference between object and partner's attractive level
diffSinc	difference between object and partner's sincere level