CatBoost and Logistic Regression as Machine Learning Approaches in Matchmaking and Perceived Availability

Jorge Iván Pincay-Ponce
Universidad Laica Eloy Alfaro de Manabí, ULEAM
jorge.pincay@uleam.edu.ec
Manta, Manabí, Ecuador

María Roxana Martínez
Universidad Abierta Interamericana, UAI
roxana.martinez@uai.edu.ar
CABA, Buenos Aires, Argentina

Wilian Richart Delgado-Muentes
Universidad Laica Eloy Alfaro de Manabí, ULEAM
wilian.delgado@uleam.edu.ec
Manta, Manabí, Ecuador

Juan Alberto Figueroa-Suárez
Universidad Laica Eloy Alfaro de Manabí, ULEAM
juan.figueroa@uleam.edu.ec
Manta, Manabí, Ecuador

DOI: https://doi.org/10.56124/encriptar.v7i14.009

ABSTRACT

This paper aims to redesign the analysis of the “Speed Dating” dataset, which was part of the research titled “Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment,” presented by Raymond Fisman, Sheena Iyengar, Emir Kamenica, and Itamar Simonson in The Quarterly Journal of Economics, the oldest professional journal of economics in the English language, in 2006. Based on the theory of "perceived availability," which suggests that people are more likely to find those who seem more attainable or interested in them to be attractive, logistic regression and the CatBoost ensemble method were employed to uncover patterns that appear influential in the decisions of individuals of the opposite sex regarding the potential for a future relationship from a four-minute speed dating social experiment. The findings indicate that, in general, individuals prioritize the following in their potential partners, from most to least important: attractiveness, perceived compatibility, shared interests, sense of humor, ambition, satisfaction with acquaintances (indicative of sociability), TV interests, sincerity, and partner's age. These results

report an accuracy of over 80% with Logistic Regression and 88% with the CatBoost ensemble method. The tool used in model development was Orange Data Mining 3.37.

Keywords: Matchmaking, ensemble, speed dating

CatBoost y Regresión Logística como enfoques de aprendizaje automático en el Matchmaking y la Disponibilidad Percibida

Resumen

El presente trabajo tiene como objetivo rediseñar el análisis del conjunto de datos “Speed Dating”, que fue parte de la investigación titulada “Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment” presentada por Raymond Fisman, Sheena Iyengar, Emir Kamenica y Itamar Simonson, en el The Quarterly Journal of Economics, la más antigua revista de economía en idioma inglés, en 2006. Con base en la teoría de la "disponibilidad percibida" que indica que las personas son más propensas a considerar atractivas a aquellas que parecen más alcanzables o interesadas en ellas, se ha empleado regresiones logísticas y el método de ensemble CatBoost, para generar con ellos un trabajo conjunto y descubrir patrones aparentemente influyentes en la decisión de personas del sexo opuesto sobre llevar una eventual relación de pareja a partir de un experimento social de citas rápidas de cuatro minutos. Se encontró que en general, los compañeros y compañeras de las parejas privilegian en el orden de más a menos lo siguiente en sus parejas: atractivo, probabilidad de compatibilidad, intereses comunes, divertido o divertida, ambición, satisfacción con conocidos (indicativo de asocialidad), intereses de TV, sinceridad, edad de la pareja. Estos resultados reportan una exactitud superior al 80% con Regresión Logística y 88% con el método de ensamble CatBoost. La herramienta utilizada en la elaboración del modelo fue Orange Data Mining 3.37.

Palabras clave: Emparejamiento, métodos de ensemble, citas rápidas.

1. Introduction

The beginning of a relationship involves moving from a state of indecision to deciding whether to establish a relationship. Although this is a topic relevant to everyone, there has been relatively little research on these initial moments (McFarland et al., 2024), particularly involving machine learning

algorithms. However, research from the 1990s indicated that men placed more emphasis on physical attractiveness rather than intelligence or ambition, while women placed greater emphasis on income potential, considering attributes such as ambition, intelligence, and social status, especially when choosing a partner for long-term relationships. (Buss & Schmitt, 1993; Regan, 1998).

This work aims to redesign the analysis of the 'speed dating' dataset, which was part of the research titled 'Gender Differences in Mate Selection: Evidence from a Speed Dating Experiment' by Raymond Fisman, Sheena Iyengar, Emir Kamenica, and Itamar Simonson, published in The Quarterly Journal of Economics, the oldest professional economics journal in the English language, in 2006.

Today, the dataset used by the researchers, which initially applied linear regression models and occasionally separated the analysis between the 4,194 men and 4,184 women who participated in the study, is available from various websites.

Our redesign of the original analysis, at the data preprocessing level, involves generating new categorical attributes to utilize columns with over 40% missing values and others where numerical values were used but lacked numerical significance, serving as textual response options instead.

Although the importance of class balancing in classification problems is recognized (Fernández et al., 2018), it is not employed in this research. To minimize noise in the data, the approach focuses on preventing overfitting or underfitting through LASSO regularization in the logistic regression algorithm and RIDGE regularization in the CatBoost ensemble algorithm, along with other hyperparameters.

In the original paper, the researchers used the attribute 'Decision to pursue or not pursue a relationship' as the class variable. Specifically, they analyzed the data using linear regression models, arranged 22 in-person speed

dating sessions of four minutes each, and quantified each participant's expectations regarding how many people they thought would be interested in dating them. (Fisman et al., 2006).

It is believed that these inputs have been foundational for popular online platforms that help people find suitable partners, using recommendation systems designed to assist users in discovering other individuals who might also be interested in them. (Kleinerman et al., 2018).

This research has chosen to use the classic 'speed dating' dataset due to the advantage of having social profile data of users who were physically observed. It is assumed that certain forms of communication reveal a person's relational state, whether they are indecisive, desire a relationship, or do not want one. Additionally, certain forms of communication can persuade individuals to transition between relational states, overcoming indecision and reaching a definitive relational decision. (McFarland et al., 2024).

Previous work proposing reciprocal recommendation systems for online dating websites is highly valued. Based on the theory of 'perceived availability,' it is believed that people are more likely to find attractive those who seem more attainable or interested in them. (Association for the Advancement of Artificial Intelligence, 2018; Brannan & Mohr, 2018; Sharabi & Dorrance-Hall, 2024; Ye et al., 2024; Zheng et al., 2022),

Reciprocity is a crucial factor in relationship formation. Expectations and behaviors in dating have evolved, with a growing focus on mutual understanding and emotional compatibility. This encourages the study of how mutual interest develops in these interactions. (Brannan & Mohr, 2018). This is why this research focuses on the perception of each participant's potential partner in deciding a possible match. Specifically, the target column to be analyzed is the partner's early decision on the day of the event.

Today, it is known that Machine Learning can identify patterns in data that are not apparent to humans. In the context of dating, it can detect subtle

compatibilities between users based on less obvious behaviors and characteristics. The use of Machine Learning for identifying potential partners is known as matchmaking. (Hayashi et al., 2023; McFarland et al., 2024).

This paper uses the original dataset and presents a computational approach based on Logistic Regression with Lasso Regularization. This approach is chosen due to the model's high interpretability, the implicit feature selection provided by the regularization, and the consequent clear communication of the results. (Weigard & Spencer, 2023).

Additionally, since many features in the dataset were discretized, the CatBoost classifier, based on ensemble methods, is used. CatBoost is valued for its robustness against overfitting and its ability to provide competitive performance compared to other boosting algorithms like XGBoost and LightGBM. Its innovations in handling categorical variables significantly reduce the need for traditional techniques such as one-hot encoding, which can increase problem dimensionality and training time. (Joshi et al., 2021; Pincay Ponce et al., 2024; Prokhorenkova et al., 2019).

2. Methodology

2.1 Data Preparation

This section documents the changes made in the data preparation beyond those performed on the original "speed dating" dataset (Fisman et al., 2006):

- **Discretization of Numerical Features:** Numerical features were discretized due to the use of linear regression models to address the problem (Fisman et al., 2006). High numbers did not imply better or worse but merely different options, which could introduce noise into the results (Pincay-Ponce et al., 2020). Features discretized include gender, race, their goal on a date, frequency of dates, frequency of outings, field of study, the

quality-price ratio of the school they graduated from, average family income, and whether it matters if their partner is of the same race, among others.

- **Omission of 1-10 Rating Features:** Features rated on a 1-10 scale, where 1 is terrible and 10 is excellent, were omitted. These ratings reflect what each person believes is most important to members of the opposite sex when deciding to date someone, concerning six key characteristics: attractiveness, sincerity, intelligence, charm, ambition, and shared interests. The feature concerning the number of past dates was also omitted because 92% of participants did not provide this information.
- **Retention and Omission of 1-10 Scale Features**: Features on a 1-10 scale were retained, while equivalent scales in 1-100 were omitted. Retained features include beliefs about how their partner (regardless of gender) perceives them, how they think men and women perceive the six attributes, how they think their date perceives them, their self-perception, their rating of a yes or no during the experimental date, satisfaction with known people, and their self-perception at the beginning of the date.
- **Scaling of Perception Features**: Features related to the perceived qualities of what their partner might have, and the general perceptions of other men and women were scaled to integer ranges of 1-10, as they were originally in 1-100. This allows discretization into only ten possible values.
- **Scaling of Match Expectation Features**: Features regarding match expectations were scaled to integer ranges of 1-10, from an original range of 1-20, thus discretizing them into only ten possible values.
- **Imputation of Missing Data**: Columns with missing data that were not omitted from the analysis were imputed with the median for numerical data or the mode for categorical data, as these imputation options are less sensitive to outliers (Pincay Ponce et al., 2024).
- **Creation of Categorical Features**: A categorical feature was created to identify the initial correlation of interests in a potential partner, categorized

as highly similar (correlation greater than 0.6), highly different (correlation less than 0.6), or of low relevance in other cases.

- **Creation of Field of Study Correspondence Feature**: A feature was created to identify the match between a person's field of study and their desired career path.

2.1 Regarding data modeling with logistic regression

The models, including data preprocessing, were developed in Orange Datamining software. 8378 instances (100%) and 97 features of 187 originally available were considered for analysis, even so the resulting data set was of high dimensionality, so regularized Logistic Regression with Least Absolute Shrinkage and Selection Operator (LASSO, L1) was used, thus generating a selection of features by setting Odds to zero (Pincay Ponce, 2023), and, consequently, setting to zero the possibility that certain pairs of characteristic-value combinations affect the YES/NO decision to make an appointment on the day of the event (Pincay Ponce, 2023).

The logistic regression is modeled as $\hat{p} = \frac{1}{1+\mathcal{E}^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 \ldots+\beta_n x_n)}}$, where x is the vector of the 96 predictor features, $\beta_1 x_1, \beta_1 x_1, \beta_2 x_2, \beta_3 x_3 \ldots$ is the dot product of each feature vector by the model coefficients β, β_0 is the bias terms, and e is Euler's irrational numerical constant (approximately equal to 2.71828). To train such a model, the loss function based on the negative likelihood function is minimized:

$$Loos\ (w, b) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log\left(p(x_i)\right) + (1 - y_i) log\left(1 - p(x_i)\right)\right] \qquad (1)$$

Where w represents the vector of coefficients or weights associated with each of the features x_i of the model, N is the number of instances, y_i is the actual output or class for observation x_i. Each component w_i corresponds to the weight or coefficient that multiplies feature x_{ij} at the i-th instance. To avoid

overfitting or underfitting and force feature selection by causing some coefficients to be exactly zero, an L1 or LASSO penalty term is added to the loss function $\lambda\|w\|_1$, so the model training becomes:

$$Loos\ (w,b) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log\left(p(x_i)\right) + (1-y_i) log\left(1-p(x_i)\right)\right] + \lambda\|w\|_1 \quad (2)$$

Where λ is the regularization parameter that controls the magnitude of the penalty, and $\|w\|_1 = \sum_{j=1}^{d}|w_j|$ is the L1 regularization norm, LASSO. The LOSS of the logistic regression coefficients were calculated from three approaches (a) Decision of all the participants' partners to match or not on the day of the event, (b) Decision of the participants' male partners and (c) Decision of the participating male partners. After several trials and errors, in the search for a balance between the capacity of accuracy, generalization and simplicity of the logistic regressions, the following hyperparameters were established:

Table 1. Logistic regressions, regularization hyperparameters and percentage of accuracy achieved.

Subject of analysis	Regularization	λ	*Accuracy %*
All participants	L1, LASSO	0.003	*81.6*
Female peers	L1, LASSO	0.007	*85.0*
Male peers	L1, LASSO	0.006	*80.0*

Source: Investigation

With regularization strength values λ between 0.003 y 0.007 it was possible to avoid both extreme overfitting and underfitting, give a good treatment to the problem of class imbalance and provide a balance between the prediction capacity and the simplicity of the model. It was detected that values greater than this achieved a Classification Accuracy of 100%, indicative of a model without generalization capacity and that is capturing trivial patterns.. (Pincay Ponce, 2023).

Now, the Odds Ratio, $\mathcal{E}^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 \ldots+\beta_n x_n)}$, is used to interpret the coefficients in terms of odds, using the exponential of the coefficients. This value, known as the odds ratio, indicates how much the event's odds change

for a unit increase in x_i. For example, if $\beta_i = 0.5$, then $\mathcal{E}^{-0.5} \approx 1.65$, which means that for each additional unit in x_i the odds of the event occurring are multiplied by approximately 1.65. So, to make the results more interpretable, the log-odds are expressed in probabilities by the following formula:

$$p = \frac{1}{1+e^{-z}} \qquad (3)$$

Where -z is the negative of each log-odd of the coefficients and e is Euler's irrational numerical constant (2.71828).

2.2 Regarding data modeling with CatBoost

CatBoost is a gradient boosting algorithm that is particularly effective at handling data sets with categorical features, as is the case for most features after they have been preprocessed for this research. (Pincay Ponce et al., 2024; Prokhorenkova et al., 2019). Furthermore, it differs from other gradient boosting such as XGBoost and LightGBM because it builds balanced trees that are symmetric in structure, this means that at each step, the same split feature pair that results in the lowest loss is chosen and applied to all nodes at that level. The hyperparameters considered are:

Table 2. CatBoost and regularization hyperparameters.

HYPERPARAMETER	VALUE	ROLE IN THE ALGORITHM
Number of Trees, NNN	100	Controls the model's complexity and its capacity to fit the data.
Learning Rate Determines the size at which the weights of the n trees are updated.	0.2	0.2 is a high learning rate, which can accelerate training but may increase the risk of overfitting; however, in this case, overfitting was not present.
Reproducible Training	True	Ensures that the training results are reproducible.
Regularization $\boldsymbol{\lambda}$ L2, RIDGE.	0.19	0.19 is a low value that does not address overfitting but rather underfitting, which is relevant to this research.
Maximum Tree Depth	5	5 is a moderate value that balances accuracy and manages overfitting or underfitting.
Feature Subset	0.5	A medium value (50%) that reduces feature correlation and improves feature eligibility.

Source: Investigation

In the context of CatBoost, Ridge regularization $\lambda \sum_{j=1}^{n} w_j^2$ is mathematically expressed as an additional term in the loss function:

$$Loos\ (w) = L(y, \hat{y}) + \lambda \sum_{j=1}^{n} w_j^2 \qquad (4)$$

Where $L(y, \hat{y})$ is the primary loss function, in the form of Log-Odds, for the classification problem presented here, λ is the regularization parameter that acts as the regularization strength, which in this case is minimal. w_j^2 represents the square of each CatBoost coefficient, and $\sum_{j=1}^{n} w_j^2$ is the sum of the squares of all model coefficients.

To facilitate the interpretation of CatBoost, SHAP values, which stands for Shapley Additive Explanations, were used. These address the issue that models often produce output values that are not easily interpretable because they focus on the 'how much?' of the problem, meaning the reasons behind the outputs are unknown. SHAP values illustrate how each feature affects the prediction, even for complex methods such as gradient boosting (as with CatBoost) or neural networks.(Lundberg, 2018; Van den Broeck et al., 2022).

3. Resulted

Based on the feature selection provided by LASSO regularization in Logistic Regression, we present the probabilities, from highest to lowest, of the influence of feature-value pairs on the decision on the first day of the event by female partners and male partners, regarding whether they will eventually match, i.e., that first impression. Table 3 shows the alignment in preferences between the partner's view of the participant's interests and what they consider attractive. Female partners valued their interest in concerts, TV, music, sports, and the age of both parties. Male partners valued the age of the women, but not their own

Table 3. Influence of features – value on the probability that partners decide to match.

	Women's companions	P.%	Companions of men	P.%
1	RatingByPartnerLikeYou	0.61	RatingByPartnerLikeYou	0.57
2	RatingByPartnerAttractive	0.59	RatingByPartnerAttractive	0.55
3	RatingByPartnerProbalityMatch	0.53	RatingByPartnerSahredInterest	0.52
4	RatingByPartnerSahredInterest	0.50	RatingByPartnerFun	0.52
5	YourInterestingConcerts	0.50	RatingByPartnerProbalityMatch	0.51
6	YourInterestingTv	0.50	YourYesNoDuringDatingAttractive	0.51
7	YourInterestingMusic	0.50	ProbableThatPersonAccept	0.50
8	YourInterestingTVsports	0.50	Self-perceptionPartnerFunTime1	0.50
9	Self-perceptionPartnerSincereTime1	0.50	Self-perceptionPartnerIntelligentTime1	0.50
10	Self-perceptionPartnerSharedInterestTime1	0.50	YourAge	0.50
11	Self-perceptionPartnerAmbitiousTime1	0.50	Self-perceptionPartnerSincereTime1	0.49
12	Self-perceptionPartnerIntelligentTime1	0.49	Self-perceptionPartnerAmbitiousTime1	0.49
13	AgePartner	0.49	Self-perceptionPartnerAttractiveTime1	0.49
14	YourAge	0.49	AttractivePartner	0.46
15	Self-perceptionPartnerAttractiveTime1	0.49	match=No	0.13
16	Self-perceptionPartnerFunTime1	0.49		
17	YourInterestingTheater	0.49		
18	AttractivePartner	0.48		
19	RatingByPartnerSincere	0.48		
20	match=No	0.29		

We present the probabilities, from highest to lowest, of the influence of feature-value pairs on the partners, without distinction of gender.

Table 4. Influence of features – value on the overall probability that partners decide to match.

	All	P.%
1	RatingByPartnerAttractive	0.57
2	RatingByPartnerProbalityMatch	0.52
3	RatingByPartnerSahredInterest	0.51
4	RatingByPartnerFun	0.50
5	RatingByPartnerAmbitious	0.50
6	SatisfactionWithAcquaintances	0.50
7	YourInterestingTVsports	0.50
8	RatingByPartnerSincere	0.50
9	AgePartner	0.50
10	Self-perceptionPartnerSincereTime1	0.50
11	Self-perceptionPartnerIntelligentTime1	0.50
12	Self-perceptionPartnerFunTime1	0.49
13	Self-perceptionPartnerAttractiveTime1	0.49

14	YourAge	0.49
15	Self-perceptionPartnerAmbitiousTime1	0.49
16	AttractivePartner	0.47
17	match=No	0.21

The confusion matrix in **Figure** 1 shows the accuracy results for each class. The most used metric in a classification problem like this is accuracy. For illustration, TP stands for true positive, referring to the set of instances for which the model's prediction was correct. True negative (TN) refers to the set of instances for which the prediction was correct; false positive (FP) is the number of instances where the prediction was incorrect, and false negative (FN) is the number of instances where the prediction was incorrect.

Therefore, accuracy is calculated as $\text{Accuracy} = \frac{TP_1+TP_2+\cdots+TP_n}{TP+FP}$ (Mukhopadhyay, 2018). Precision is a measure of correctness that indicates how many of the predicted positive instances are positive. It is calculated as $\text{Precision} = \frac{TP}{TP+FP}$ (Mukhopadhyay, 2018). Recall is a measure of how many actual positive observations are correctly predicted; it is also known as sensitivity. It is calculated as $\text{Recall} = \frac{TP}{TP+FN}$ $(\text{Mukhopadhyay}, 2018)$.

Figure 1. Accuracy results for each class according to the CatBoost algorithm.

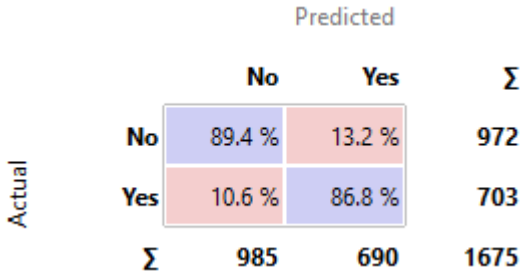
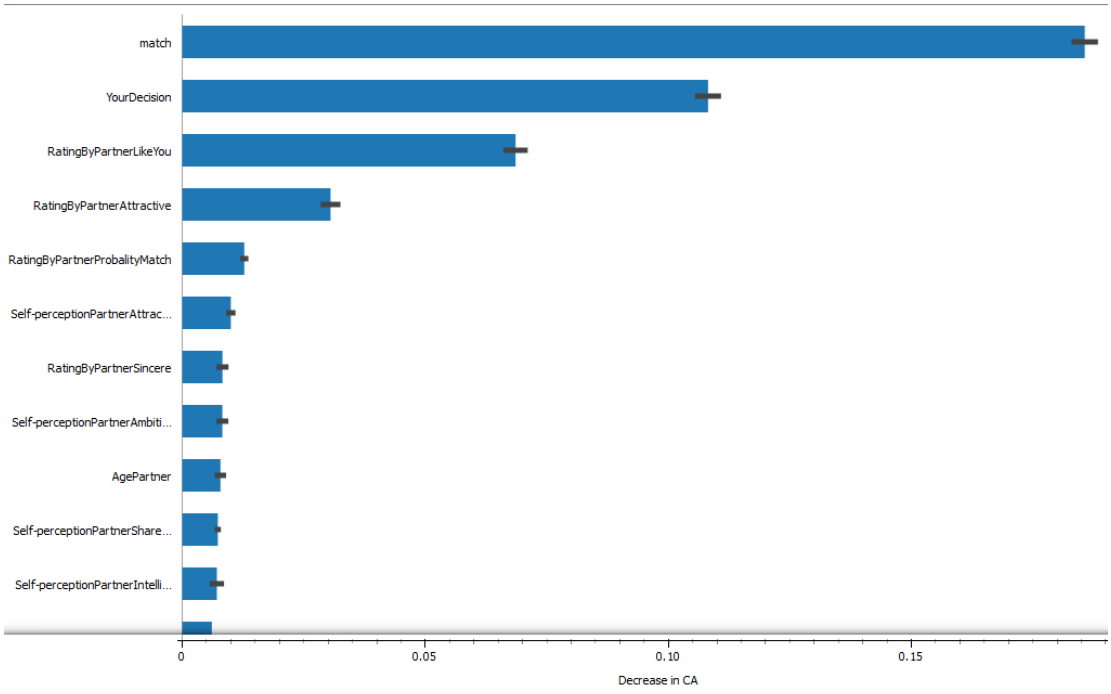


Table 5. Accuracy, precision, recall, and obtained times reported by all models. The period is used as the decimal separator.

Model	CA	Prec	Recall
CatBoost. RIDGE	0,88	0,88	0,88
Logistic Regression. Male. LASSO	0,85	0,85	0,85
Logistic Regression. ALL	0,82	0,82	0,82
Logistic Regression. Female. LASSO	0,80	0,80	0,80

The interpretation of CatBoost has benefited from the SHAP framework proposed in 2016 by Scott Lundberg and Su-In Lee (2017), because SHAP proved to be an easy and theoretically sound way to understand the predictions of any model. Figure 2 illustrates these details, but it does not specify the feature-value pairs, which can often be inferred from the logistic regression results.

Figure 2. Influence of features on the classification performed by the CatBoost algorithm.



4. Conclusions

There is still much to learn about attraction and relationship formation, but the innovative contribution provided by machine learning is valuable for exploring dyadic behavior, which can offer new empirical insights into how people choose partners and form relationships. We have introduced a foundation for generating new research questions on first impression formation, romantic rivalries, and affiliative behaviors. Here, we highlight the use of data, such as colleges with a high Quality-Price ratio for tuition, which is valued numerically between 800 and 1600—the best value—as it provides an indicator of each participant's economic expectations.

Following this, in general, individuals prioritize the following traits in their partners, from most to least important: attractiveness, perceived compatibility, shared interests, sense of humor, ambition, satisfaction with acquaintances (indicative of sociability), TV interests, sincerity, and the partner's age. Additionally, the first impression that the other person projects in terms of sincerity, intelligence, grace, and attractiveness is also considered. Subtle gender differences, which were presented in the results section, do exist.

It is acknowledged, but not studied in this research, that on dating platforms, empirical evidence suggests that the likelihood of a user being recommended by the platform's algorithm increases significantly, and perhaps biasedly, with the user's popularity. Therefore, the analysis presented by us is more traditional or conservative.

It is crucial to acknowledge the limitations of this study and its specific context. The data originate from a speed dating experiment, which may not fully reflect partner selection processes in other environments. Furthermore, technological advancements and social changes since the original data collection may have influenced dating preferences and behaviors. Future studies should consider replicating this analysis with more recent and diverse datasets, as well as exploring how algorithms in online dating platforms

influence matching decisions. Additionally, investigating how preferences and behaviors vary across different cultures and demographic groups would be valuable. These considerations are particularly relevant in the context of machine learning, as the quality and representativeness of training data significantly impact model performance and generalizability. Moreover, the dynamic nature of human behavior and societal norms underscores the importance of continually updating our models and methodologies to ensure they remain relevant and accurate in capturing the complexities of human mate selection processes.

5. Future research

Future research directions should encompass a multifaceted approach to understanding the complexities of mate selection in the modern era. Firstly, exploring how perceptions of availability evolve over time in long-term relationships could provide valuable insights into relationship dynamics. Secondly, investigating the impact of social media and dating applications on partner selection strategies is crucial in our increasingly digital world. Thirdly, analyzing how cultural factors influence the prioritization of attributes in mate selection would offer a more comprehensive, global perspective on this phenomenon. Fourthly, studying the influence of artificial intelligence and recommendation algorithms in online dating platforms is essential to understand the technological mediation of romantic connections. Lastly, examining how the theory of "perceived availability" interacts with other psychological models of attraction could lead to a more integrated understanding of human mating behaviors. These diverse research avenues would significantly contribute to our knowledge of contemporary mate selection processes and their societal implications.

6. References

Association for the Advancement of Artificial Intelligence (Ed.). (2018). *Proceedings of the Twelfth International AAAI Conference on Web and Social Media: ICWSM: 25-28 June 2018, Stanford, California, USA*. International AAAI Conference on Web and Social Media, Palo Alto, California. AAAI Press.

Brannan, D., & Mohr, C. D. (2018). Love, friendship, and social support. *Noba textbook series: Psychology. Champaign, IL: DEF publishers*.

Buss, D. M., & Schmitt, D. P. (1993). Sexual Strategies Theory: An evolutionary perspective on human mating. *Psychological Review*, *100*(2), 204-232. https://doi.org/10.1037/0033-295X.100.2.204

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. https://doi.org/10.1162/qjec.2006.121.2.673

Hayashi, T., Mawalim, C. O., Ishii, R., Morikawa, A., Fukayama, A., Nakamura, T., & Okada, S. (2023). A Ranking Model for Evaluation of Conversation Partners Based on Rapport Levels. *IEEE Access*, *11*, 73024-73035. https://doi.org/10.1109/ACCESS.2023.3287984

Joshi, A., Saggar, P., Jain, R., Sharma, M., Gupta, D., & Khanna, A. (2021). CatBoost—An Ensemble Machine Learning Model for Prediction and Classification of Student Academic Performance. *Advances in Data Science and Adaptive Analysis*, *13*(03n04), Article 03n04. https://doi.org/10.1142/S2424922X21410023

Kleinerman, A., Rosenfeld, A., Ricci, F., & Kraus, S. (2018). Optimally balancing receiver and recommended users' importance in reciprocal recommender systems. *Proceedings*

of the 12th ACM Conference on Recommender Systems, 131-139. https://doi.org/10.1145/3240323.3240349

Lundberg, S. (2018). *SHAP*. API Reference. https://tinyurl.com/yhcdt2w8

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. https://doi.org/10.48550/ARXIV.1705.07874

McFarland, D. A., Broska, D., Prabhakaran, V., & Jurafsky, D. (2024). Coming into relations: How communication reveals and persuades relational decisions. *Social Networks*, *79*, 57-75. https://doi.org/10.1016/j.socnet.2024.05.003

Mukhopadhyay, S. (2018). *Advanced Data Analytics Using Python*. Apress. https://doi.org/10.1007/978-1-4842-3450-1

Pincay Ponce, J. I. (2023). *Análisis de datos educativos aplicado en el estudio de la incidencia de factores socioeconómicos en el rendimiento escolar* [Doctor en Ciencias Informáticas, Universidad Nacional de La Plata]. https://doi.org/10.35537/10915/156471

Pincay Ponce, J. I., De Giusti, A. E., Sánchez Andrade, D. A., & Figueroa Suárez, J. A. (2024). CatBoost: Aprendizaje automático de conjunto para la analítica de los factores socioeconómicos que inciden en el rendimiento escolar. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, *38*, e3. https://doi.org/10.24215/18509959.38.e3

Pincay-Ponce, J., Sánchez-Andrade, D., Caicedo-Ávila, I., & Macías-Valencia, D. (2020, noviembre 27). *Clasificación de pacientes según su posibilidad de adquirir Diabetes Mellitus empleando algoritmos de Machine Learning*. IV Congreso Internacional Tecnologías de la Información y Computación (CITIC 2020), Calceta, Ecuador. https://tinyurl.com/yve333v7

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2019). *CatBoost: Unbiased boosting with categorical features* (arXiv:1706.09516; Número arXiv:1706.09516). arXiv. http://arxiv.org/abs/1706.09516

Regan, P. C. (1998). Minimum Mate Selection Standards as a Function of Perceived Mate Value, Relationship Context, and Gender. *Journal of Psychology & Human Sexuality*, *10*(1), 53-73. https://doi.org/10.1300/J056v10n01_04

Sharabi, L. L., & Dorrance-Hall, E. (2024). The online dating effect: Where a couple meets predicts the quality of their marriage. *Computers in Human Behavior*, *150*, 107973. https://doi.org/10.1016/j.chb.2023.107973

Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the Tractability of SHAP Explanations. *Journal of Artificial Intelligence Research*, *74*, 851-886. https://doi.org/10.1613/jair.1.13283

Weigard, A., & Spencer, R. J. (2023). Benefits and challenges of using logistic regression to assess neuropsychological performance validity: Evidence from a simulation study. *The Clinical Neuropsychologist*, *37*(1), 34-59. https://doi.org/10.1080/13854046.2021.2023650

Ye, Y., Ni, K., Jing, F., Zhou, Y., Tang, W., & Zhang, Q. (2024). Model-Informed Targeted Network Interventions on Social Networks Among Men Who Have Sex With Men in Zhuhai, China. *IEEE Transactions on Computational Social Systems*, *11*(1), 238-246. https://doi.org/10.1109/TCSS.2022.3216756

Zheng, X., Zhao, G., Zhu, L., Zhu, J., & Qian, X. (2022). What You Like, What I Am: Online Dating Recommendation via Matching Individual Preferences with Features. *IEEE Transactions on Knowledge and Data Engineering*, 1-1. https://doi.org/10.1109/TKDE.2022.3148485