Analysis of Factors Influencing Love Confession and Result Prediction Based on Artificial Intelligence Algorithms and Questionnaire Surveys

Yanming Chen[*], Xiaolin He[a]

Shantou University, Shantou, Guangdong, China

[*]21ymchen@stu.edu.cn; [a]23xlhe@stu.edu.cn
(The two authors make the same contribution to this paper)

Abstract. This pioneering research integrates human-centric data science into the field of psychology, specifically focusing on the analysis of factors influencing love confession and outcome prediction. By conducting a comprehensive questionnaire survey, we collected love confession-related data from 1000 young individuals in China. Following preprocessing techniques such as normalization, SMOTE oversampling, and polynomial feature engineering, we conducted in-depth analysis of the influencing factors and their correlations using methods including Point-biserial analysis, Spearman correlation coefficient, and random forest feature importance. Subsequently, we employed various machine learning algorithms, such as logistic regression, decision tree, randomforest, xgboost, Adaboosting and Stacking, to construct classification models for predicting love confession outcomes, achieving a maximum F1-score of 0.9 on the test dataset. Finally, we successfully employed a feed-forward neural network model, achieving an F1-score of 1.0 for love confession outcome prediction. This study represents a successful exploration of the novel application of artificial intelligence algorithms in the context of love confession.

Keywords: Love Confession; Artificial Intelligence; Machine learning; Questionnaire Surveys; Human-Centric Data Science

1 Introduction

Human-centered data science is an emerging interdisciplinary field that focuses on utilizing data analytics techniques and advanced technologies to gain deep insights into human behavior, preferences, and experiences. It places human factors at the core of data-driven research and applications, aiming to understand and address human needs and challenges[1]. The core of human-centered data science combines principles from various disciplines, including computer science, statistics, psychology, and sociology[2].

D. Rad et al. (eds.), *Proceedings of the 2024 5th International Conference on Mental Health, Education and Human Development (MHEHD 2024)*, Advances in Social Science, Education and Humanities Research 857,
https://doi.org/10.2991/978-2-38476-271-2_79

Love confession, as a cross-disciplinary field of human-centered data science and psychology, has been rarely predicted through algorithms in previous research. For instance, previous studies such as "Mathematical Model of Male Pursuing Female" utilized differential equations to derive pursuit strategies[3]. However, previous research has been unable to effectively predict love confession outcomes, and the conclusions obtained have not been well applicable in practice.

This paper introduces, for the first time, a modeling and prediction of love confession outcomes using a combination of artificial intelligence algorithms and questionnaire surveys, opening up new avenues for exploring and analyzing complex love confession behaviors. In this study, we collected comprehensive data through a questionnaire survey of 1000 young individuals in China. After conducting feature engineering and analyzing the correlation of influencing factors, we employed machine learning algorithms, including logistic regression, decision tree, randomforest, xgboost, Adaboosting, and Stacking, to construct classification models for predicting love confession outcomes[4]. Additionally, we utilized a feed-forward neural network model to achieve more accurate predictions, ultimately achieving accuracy and F1-score of 1.0 on the test set, demonstrating the excellent performance of the model.

This study can alleviate some uncertainties and anxieties, helping individuals make more informed decisions. When faced with the decision to love confess, individuals can utilize the predictive results of this study as a reference, assisting them in evaluating and analyzing potential confession outcomes from a scientific perspective. Furthermore, it provides valuable insights for researchers and practitioners in psychology and relationship counseling, aiding in better understanding human behavior and psychological processes and providing scientific support for related fields' theories and practices.

2 Method

2.1 Dataset Used in the Study

The love confession-related data used in this study consisted of responses obtained through a questionnaire survey from 1000 young individuals in China. Some questionnaire surveys tend to inquire about the subjective feelings and thoughts of the respondents. However, this subjectivity can lead to inaccuracies in predicting outcomes. Therefore, this study primarily focused on collecting objective information variables to mitigate the subjective nature of the questionnaire. The dataset comprised 21 columns, including variables such as love confession outcome (1 for success, 0 for failure), gender, age, height, weight, educational gap, duration of acquaintance, different location, duration of pursuit, frequency of communication, frequency of hanging out, and so on. The love confession outcome served as the dependent variable, while the remaining 20 variables were treated as independent variables. Among the 1000 observations, 571 love confessions were successful, while 429 were unsuccessful.

To begin, we transformed the categorical variables in the dataset. Categorical variables with an ordinal relationship were assigned numerical values, while other categorical variables were converted into dummy variables using one-hot encoding in subsequent analyses[5]. For instance, variables such as "self-appearance level" and "appearance level of the confession target" had four ordinal categories: low, moderate, high, and very high. We defined these categories as 0, 1, 2, and 3, respectively. Similarly, the variables "frequency of hanging out" and "frequency of communication" also had four categories: never, rarely, occasionally, and frequently, which we defined as 0, 1, 2, and 3, respectively. The same approach was applied to other variables. After the conversion, we obtained the basic descriptive statistics for the entire dataset, as presented in Table 1.

Table 1. Rudimentary descriptive statistics of the numerical variables and categorical variables

	Min	Mean	Max	Std	Median	Count
Love confession outcome	0.00	0.54	1.00	0.50	1.00	1000.00
Gender	0.00	0.45	35.00	0.50	0.00	1000.00
Age	13.00	21.26	35.00	3.87	21.00	1000.00
Age of the confession target	13.00	21.36	189.00	3.79	21.00	1000.00
Height	147.00	167.93	92.00	7.88	168.00	1000.00
Weight	39.90	57.65	3.00	9.71	56.00	1000.00
Self-appearance level	0.00	1.47	3.00	0.61	1.00	1000.00
Appearance level of the confession target	0.00	1.92	1.00	0.69	2.00	1000.00
Educational gap	0.00	0.06	1.00	0.24	0.00	1000.00
Different location	0.00	0.25	1.00	0.43	0.00	1000.00
Duration of acquaintance	0.00	2.14	4.00	1.21	2.00	1000.00
Duration of pursuit	0.00	2.15	4.00	1.27	2.00	1000.00
Whether they were close friends	0.00	0.58	1.00	0.49	1.00	1000.00
Frequency of hanging out	0.00	1.90	3.00	0.95	2.00	1000.00
Frequency of communication	0.00	2.61	3.00	0.66	3.00	1000.00
Compatibility of hobbies	0.00	2.18	4.00	0.97	2.00	1000.00
	Unique	Mode	Top Frequency	Count		
Educational background at the time of confession	5.00	Undergraduate	608	1000.00		

Personality traits	3.00	Extroverted personality	419	1000.00
Method of acquaintance	6.00	Classmate/Colleague relationship	656	1000.00
Attitude of the confession target when replying to messages online	4.00	Normal	432	1000.00
Form of love confession	4.00	Face to face	419	1000.00

Dummy variables, also referred to as indicator variables, are binary variables employed to represent categorical variables[6]. In our study, we employed one-hot encoding to transform the five categorical variables in Table 1 into dummy variables, facilitating subsequent analysis.

2.2 Exploration of Correlation

The Point-Biserial Correlation is a statistical method used to measure the correlation between a binary variable and a continuous variable[7]. It is calculated based on the Point-Biserial Correlation Coefficient, which ranges from -1 to 1. In Point-Biserial correlation analysis, the p-value is used for significance testing. The p-value ranges from 0 to 1, and a p-value < 0.05 indicates significant correlation, with values closer to 0 indicating higher significance. On the other hand, the Spearman correlation coefficient is a non-parametric statistical method used to measure the monotonic relationship between two variables. The Spearman correlation coefficient is calculated based on the ranks of the variables, rather than the actual values . It also ranges from -1 to 1. Additionally, in this study, the feature importance of the random forest was evaluated by calculating the decrease in Gini impurity for each feature during node splitting, using the Gini impurity-based method[8]. This quantifies the importance of each feature in predicting the outcome.

Firstly, we explored the relationships among the independent variables. By calculating the Spearman correlation coefficients between the independent variables, we observed that the majority of variables did not exhibit strong correlations. However, it is noteworthy that the variables "Frequency of communication," "Compatibility of hobbies," and "Whether they were close friends" showed some positive correlations with the variable "Frequency of hanging out," with Spearman correlation coefficients of 0.45, 0.33, and 0.36, respectively. Additionally, the variable "Positive attitude of the confession target when replying to messages online" displayed a moderate correlation with "Frequency of hanging out," with a coefficient of 0.28. Furthermore, it had a correlation coefficient of 0.24 with the variable "Self-appearance level".

Secondly, we explored the impact of each independent variable on the dependent variable using Point-Biserial correlation analysis, Spearman correlation coefficient, and random forest feature importance. The variables with high correlation are shown in Table 2.

Table 2. The variables that exhibit a high correlation with the love confession outcome

	P-value	Point-Biserial correlation coefficient	Spearman correlation coefficient	Randomforest feature importance
Positive attitude of the confession target when replying to messages online	0.000	0.395	0.396	0.174
Frequency of hanging out	0.000	0.302	0.303	0.162
Normal attitude of the confession target when replying to messages online	0.000	-0.261	-0.261	0.079
Face to face	0.000	0.234	0.235	0.075
Message (e.g., Wechat)	0.000	-0.229	-0.228	0.061
Self-appearance level	0.000	0.219	0.227	0.059
Compatibility of hobbies	0.000	0.217	0.216	0.058
Introverted personality	0.000	-0.210	-0.211	0.056
Frequency of communication	0.000	0.203	0.222	0.044
Negative attitude of the confession target when replying to messages online	0.000	-0.194	-0.157	0.036
Sometimes hot and sometime cold	0.001	-0.166	-0.144	0.032

Through the analysis of 1000 respondents, it can be concluded that features such as gender, height, weight, educational gap, and whether they are in different locations have little correlation with the love confession outcome. It is noteworthy that the duration of acquaintance and the duration of pursuit also do not show significant correlation with the love confession outcome. Therefore, in these 1000 samples, it is not necessarily the case that the longer the pursuit, the higher the probability of successful confession. The correlation is actually not substantial.

Furthermore, it was found that the attitude of the confession target when replying to messages online is an important factor. If the target responds with a positive attitude, it significantly increases the probability of a successful confession. On the other hand, if the target exhibits a normal, negative, or inconsistent attitude, it is negatively correlated with the love confession outcome. The reason why the negative attitude has a stronger negative correlation than the normal attitude may be due to sample bias. When the target responds with a negative attitude, people generally tend not to confess, resulting in fewer observed values.

Frequency of hanging out, frequency of communication, compatibility of hobbies, and self-appearance level are positively correlated with the love confession outcome, with frequency of hanging out being the most important. In terms of the choice of confession method, face-to-face confession is positively correlated, while message-based confession is negatively correlated with the love confession outcome. This may be because face-to-face confession better expresses individuals' emotions and sincerity, but further research is needed to investigate this aspect. Additionally, introverted personality is negatively correlated with the confession outcome to some extent.

Polynomial feature processing is a commonly used data preprocessing method aimed at expanding the original feature space to capture more nonlinear relationships among features[9]. It achieves this by expanding the combinations and powers of the original features into new features, enabling the model to better fit the nonlinear relationships in the data. In this study, the polynomial feature processing method is applied to analyze the impact of combined variables on the outcome of love confessions, in order to comprehensively consider the interactions among features. The study found that the combined effects of variables such as "Self-appearance level," "Frequency of communication," "Frequency of hanging out," and "Positive attitude of the confession target when replying to messages online" have a significant influence on the confession outcome. Particularly noteworthy are the variables "Positive attitude" and "Self-appearance level," followed by "Frequency of hanging out." Furthermore, there exists a certain positive correlation among these four variables themselves. When an individual scores high in at least three or all four of these variables, the model is highly likely to predict a successful outcome for their love confession.

Based on the analysis of correlation in this chapter, we can preliminarily derive the following strategies: 1. Do not deliberately prolong the pursuit time as it does not effectively increase the probability of success. Instead, follow your inner feelings; 2. Increase the frequency of hanging out and participate in activities that align with the target's interests and hobbies; 3. Pay attention to personal appearance and grooming; 4. Opt for face-to-face love confession methods whenever possible; 5. Take into account the target's attitude when responding to messages online, as a positive response increases the likelihood of success. By considering these strategies and leveraging the insights gained from the analysis, individuals can improve their approach to love confessions and increase the chances of a positive outcome.

2.3 Feature Engineering

Compared to Z-score standardization, normalization has the main advantage of not assuming any specific data distribution and preserving the relative relationships of the original data[10]. On the other hand, Z-score standardization transforms the data into a standard normal distribution with a mean of 0 and a standard deviation of 1. It is suitable for standardizing data to adhere to the assumption of a normal distribution. In contrast, normalization is more suitable for scenarios where data needs to be scaled to a specific range without considering the distribution shape. Therefore, in this study, the min-max scaling method was utilized to normalize the numerical variables. This transformation rescaled the data to the range of [0, 1]. The purpose was to improve the convergence speed of algorithms, eliminate dimensional differences among features, and mitigate the dominant influence of certain features.

Then, we divided the dataset into a training set and a test set in a ratio of 7:3. The training set and test set consisted of 700 and 300 observations, respectively. Next, we applied the SMOTE algorithm to the training set for oversampling. Unlike random oversampling, SMOTE generates new synthetic samples by interpolating between minority class samples, rather than simply duplicating existing samples. This approach reduces redundancy and repetition among samples. The principle of the SMOTE algorithm assumes that {A1}, {A2}, ..., {An} represent n classes of samples, and {Ak} has a small sample size that needs to be expanded. Starting with a selected sample a1 from this class, the algorithm calculates the distances between a1 and other samples in the same class, and selects the K nearest neighbors. Then, a random neighbor sample a2 is chosen from the K neighbors, and a new synthetic sample ak1 is generated as ak1 = ra1 + (1 − r)a2, where r ∈ [0, 1]. This process is repeated until the desired number of samples is obtained. Through the generation of synthetic samples, SMOTE can better preserve the diversity of minority class samples, alleviate the risk of overfitting, and improve the generalization ability of the model.

In this study, the PolynomialFeatures class from the scikit-learn library in Python was employed to perform polynomial feature processing on the original data. This transformation converted the original features into new features that included various higher-order terms and interaction terms of the original features. However, polynomial feature processing introduces additional features, thereby increasing the complexity and computational cost of the model. It can also potentially lead to issues such as feature dimension explosion and overfitting. Therefore, after polynomial feature processing, we applied PCA (Principal Component Analysis) for dimensionality reduction, reducing the feature space to 50 dimensions.

2.4 Machine Learning Model Building

After performing feature engineering, we applied various machine learning algorithms for modeling. Firstly, we used logistic regression[11]. Given a sample set $\{X_m\}$ formed by a set of n-dimensional column vectors $\{x_{in}\}$, we constructed a regression function to predict future scenarios. To begin with, given a parameter vector θ, also n-dimensional, the regression function is defined as $g_\theta(\mathrm{x}) = \frac{1}{1 \pm e^{-\theta' x}}$. This is a classification function that predicts the probability of an event occurrence, where the value of $g(\mathrm{x})$ represents the probability of a certain event being predicted. Then, we defined the loss function as (1):

$$\begin{cases} Cost(g_\theta(\mathrm{x}), \mathrm{y}) = -\log(g_\theta(\mathrm{x})) \quad y = 1 \\ Cost(g_\theta(\mathrm{x}), \mathrm{y}) = -\log(1 - g_\theta(\mathrm{x})) \quad y = 0 \end{cases} \tag{1}$$

The update formula for gradient descent is obtained by taking the partial derivative of θ, and it is given as (2):

$$\theta_\mathrm{j} := \theta_\mathrm{j} - \alpha \frac{1}{m} \sum_{i=1}^{m} (g_\theta(x_{(j)}^{(i)}) - y^{(i)})\, x_{(j)}^{(i)} \tag{2}$$

Next, we built a classification tree model based on the ID3 algorithm. We calculated the initial entropy $Ent(D) = \sum_i^c = p_i \log_2 p_i$, and then computed the information gain for all attribute conditions to determine the root node $Gain(D, a) = Ent(D) -$

$\sum_{i=1}^{n} \frac{|D_i|}{|D|}$, c represents the number of attributes and n represents the number of states in the corresponding attribute. The attribute condition with the highest information gain was chosen as the root node. We then repeated the same process for the child nodes until all attributes were partitioned.

Afterwards, we employed the randomforest classification algorithm for modeling. The random forest algorithm is an ensemble of multiple decision trees. The generation of these decision trees follows a random sampling procedure from a training set containing N samples. During tree splitting, only m features (m << M) are randomly selected from a pool of M features as the candidates for node selection. The selection criterion, in this case, remains the same using information gain. As a result, a large number of decision trees are generated, and the experimental subjects are assessed and voted upon by these trees. The prediction result is determined by the majority vote among the trees.

We also utilized the XGBoost and SVM-based Adaboosting classification models. XGBoost is a gradient boosting tree algorithm that optimizes the predictive performance of the model iteratively. In each iteration, a new decision tree is added, and the model's weights are adjusted based on the errors between the previous tree's predictions and the actual labels. Adaboosting, on the other hand, is an ensemble learning algorithm that enhances classification performance by sequentially training multiple classification trees. During the training process of each tree, the weights of the samples are adjusted based on the previous round's classification results, allowing the classifier to focus more on previously misclassified samples in the next round. Adaboosting combines the predictions of multiple classification trees with weighted contributions to obtain the final classification result.

Then, we employed the Stacking algorithm to integrate these classification models. The basic principle of this algorithm is to construct a meta-model that utilizes the predictions of the base models as inputs for the final prediction. In our study, when using the Stacking algorithm, the base models (first-level models) were selected as logistic regression, decision trees, randomforest classification, XGBoost, and Adaboosting. The meta-model (second-level model) utilized a decision tree model. In addition, we employed the random grid search method to determine the approximate ranges of the parameters for each classification model.

Feedforward neural networks are capable of handling complex nonlinear relationships, adapting to various types of inputs and outputs, and have good scalability. In this study, we finally constructed a five-layer feedforward neural network based on the LeakyReLU activation function and the Adam optimizer. L2 regularization was also incorporated to reduce the risk of overfitting. The core process is illustrated in Figure 1.

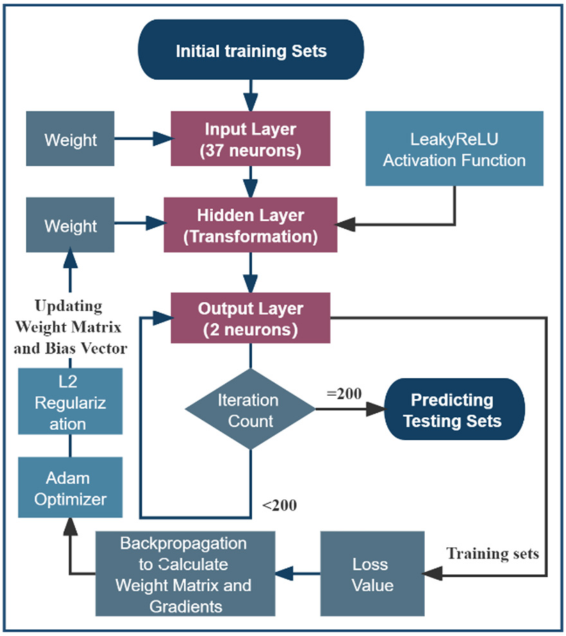

Fig. 1. The core process of the feedforward neural network constructed in this study

3 Experiments & Results

The modeling experiments conducted in this study were carried out in Python 3.8.0, and the results are shown in Table 3.

Table 3. The experiments results of different models

	Training set (Accuracy)	Testing set (Accuracy)	Testing set (F1-Score)
Logistic	0.81	0.82	0.83
DecisionTree	0.98	0.80	0.82
Randomforest	0.87	0.82	0.84
XGBoost	1.0	0.86	0.85
Adaboosting	1.0	0.83	0.83
Stacking	0.98	0.89	0.90
FNN	1.0	1.0	1.0

According to the results in Table 3, among the first five classification models, the XGBoost model has the highest training accuracy, testing accuracy, and testing F1-score. Additionally, when the five models were integrated using the Stacking algorithm, a higher F1-score of 0.9 was achieved, indicating that the Stacking algorithm effectively utilizes the strengths of different base models to improve overall predictive performance on this dataset. Furthermore, the FNN we constructed further improved the accuracy and F1-score, reaching 1.0.

4 Discussion & Conclusion

In previous studies, scholars have only obtained two conclusions: that longer pursuit time does not necessarily lead to better results and that timing is crucial. However, in our research, we have discovered that to improve the chances of a positive outcome in love confessions, individuals should avoid artificially prolonging the pursuit, trust their instincts, increase the frequency of hanging out and engage in activities aligned with the target's interests, pay attention to personal appearance and grooming, prefer face-to-face confession methods, and consider the target's attitude when responding to online messages.

Through the research presented in this paper, we have discovered that machine learning models and deep learning models can effectively predict love confession, providing a more scientifically rational basis for individual decision-making. This study also demonstrates the feasibility of applying artificial intelligence algorithms to the problem of love confessions. However, it should be noted that the observed values investigated in the experiments were limited to only 1000 instances, which may result in a lack of generalizability of the machine learning models established. This limitation could lead to biases when facing new data. Furthermore, neural network models are often considered "black box models" due to the difficulty in explaining their internal decision-making processes and prediction outcomes.

Therefore, in future research, it is crucial to expand the sample size and broaden the scope of investigation to enhance the generalizability of the models. Additionally, efforts should be made to increase the interpretability of neural network models.

References

1. Luxton, D. D. (2014). Artificial intelligence in psychological practice: Current and future applications and implications. Professional Psychology: Research and Practice, 45(5), 332.
2. Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., ... & Zhang, J. D. (2020). An introduction to machine learning. Clinical pharmacology & therapeutics, 107(4), 871-885.
3. Zhou, X., & Ke, J. Z. (2012). A mathematical model for boys chasing girls. Mathematics in Practice and Understanding, 12, 1-8.
4. Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215.
5. Al-Shehari, T., & Alsowail, R. A. (2021). An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. Entropy, 23(10), 1258.

6. Gong, X., & Lin, B. (2023). Adding dummy variables: A simple approach for improved volatility forecasting in electricity market. Journal of Management Science and Engineering, 8(2), 191-213.
7. Bonett, D. G. (2020). Point-biserial correlation: Interval estimation, hypothesis testing, meta-analysis, and sample size determination. British Journal of Mathematical and Statistical Psychology, 73, 113-144.
8. Zhao, Y., Zhu, W., Wei, P., Fang, P., Zhang, X., Yan, N., ... & Wu, Q. (2022). Classification of Zambian grasslands using random forest feature importance selection during the optimal phenological period. Ecological Indicators, 135, 108529.
9. Goyal, A., Zafar, A., Kumar, M., Bharadwaj, S., Tejas, B. K., & Malik, J. (2022). Cirrhosis Disease Classification by using Polynomial Feature and XGBoosting. In 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT),1-5. IEEE.
10. Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. Applied Soft Computing, 97, 105524.
11. Mathew, T. E., & Kumar, K. A. (2020). A logistic regression based hybrid model for breast cancer classification. Indian Journal of Computer Science and Engineering (IJCSE), 11(6), 899-903.