## **Marriage prediction data analysis report**

## **Yu Jinyan**

### **I. Introduction**

Marital quality is an important indicator of social and family stability. However, in reality, marital relations often break down due to communication barriers, lack of trust, differences in values and other problems. In the past decade, the number of divorces around the world has continued to rise, with more than 20 million couples ending their marriages every year on average.

Traditional marriage counseling relies on subjective assessments, which are inefficient and costly. Machine learning methods can uncover underlying patterns from vast amounts of behavioral data, enabling objective predictions. In modern society, analyzing behavioral patterns in marital relationships using data mining and artificial intelligence has become a new paradigm in psychology and sociology research.

The practical value of this study is reflected in two aspects:

Early warning: identify high-risk couples through behavioral characteristics to provide data support for psychological counseling;

Intervention guidance: Feature importance analysis can reveal key contradictions and help develop targeted improvement plans.

Based on a data set of 170 samples and 54 behavioral problem scores, this report uses machine learning methods to predict whether marriage is likely to end in divorce, with the goal of "predicting marital stability through behavioral data".

### **2. Problem definition**

project:

We want to build a classification model that determines whether a couple is likely to divorce. Given a set of scores for various behaviors in the relationship (0-4, from "strongly disagree" to "strongly agree"), the model outputs whether or not the couple is divorced.

Data set description:

Source: https://archive.ics.uci.edu/dataset/539/divorce+predictors+data+set

There are 170 samples and 54 behavioral characteristics, and the last column is Class, which marks whether the sample is divorced or not (1 means divorced, 0 means not divorced). The proportion of divorced samples is about 49.4% (84/170), so there is no need to deal with the problem of category imbalance.

54 features described:

1. When our conversation goes in a bad direction, one of us takes the initiative to apologize and the problem doesn't get bigger.

2. I know we can ignore our differences, even when things get tricky.

3. When we need it, we can discuss and correct with our wife from the beginning.

4. When I was having a fight with my wife, contacting him would eventually work.

5. The time I spend with my wife is special to us.

6. As partners, we don't have time to stay at home.

7. We are like two strangers sharing the same environment at home, not in a family.

8. I like to go on holiday with my wife.

9. I like to travel with my wife.

10. My goals are mostly the same as my wife's.

11. I think that one day in the future, when I look back on my past, I will find that my wife and I are in harmony.

12. My wife and I share similar values regarding personal freedom.

13. My husband and I have similar ways of entertaining ourselves.

14. We have the same goals for most people (children, friends, etc.).

15. Our dream of living with our wives is similar and harmonious.

16. We share the same view of love with our wives.

17. We share the same view with our wives about how to live a happy life.

18. My wife and I share similar views on how marriage should develop.

19. My wife and I share similar views on the distribution of roles in marriage.

20. My wife and I share similar values in terms of trust.

21. I know my wife's preferences very well.

22. I know what kind of care my wife wants when she is ill.

23. I know my wife's favorite food.

24. I can tell you what kind of pressure my wife faces in life.

25. I understand my wife's inner world.

26. I know the basic concerns of a wife.

27. I know what the source of stress is for my wife at the moment.

28. I know the hopes and wishes of my wife.

29. I know my wife very well.

30. I know my wife's friends and their social connections.

31. I can be very aggressive when I argue with my wife.

32. When I talk to my wife, I tend to use phrases like "you always" or "you never."

33. In the course of the discussion, I will use negative comments about my wife &apos; s character.

34. I use offensive expressions in my discussions.

35. I can insult our discussion.

36. I would be embarrassed when we quarrel.

37. My quarrel with my wife is not quiet.

38. I hate the way my wife brings it up.

39. Arguments often occur suddenly.

40. We started arguing before I knew what was going on.

41. When I talk to my wife about something, my composure is suddenly broken.

42. When I quarrel with my wife, my calm only comes back suddenly and I don't say a word.

43. What I most desire is for the environment to calm down a little.

44. Sometimes I feel that it is good for me to be away from home for a while.

45. I would rather remain silent than quarrel with my wife.

46. Even if I am right in the argument, I do not want to upset the other side.

47. When I quarrel with my wife, I remain silent because I am afraid that I will not be able to control my anger.

48. I felt right in our discussion.

49. I have nothing to do with the allegations against me.

50. I am not actually a person who feels guilty about the accusations against me.

51. I am not the wrong person for family issues.

52. I would not hesitate to tell her about my wife's shortcomings.

53. When I discuss this issue, I remind her of the shortcomings of my wife.

54. I am not afraid to tell her about my wife's incompetence.

### **3. Modeling method**

In order to compare the performance of different algorithms in the divorce prediction task, I selected the following three models:

1. decision tree (CART)

* Simple and intuitive, easy to understand and explain.
* Nonlinear relationships and feature interactions can be captured.
* Training is fast.

1. random forest (RF)

* An ensemble model composed of multiple decision trees.
* Reduce overfitting by voting and maintain stable performance.
* It is suitable for all types of data and is robust.

1. support vector machine (SVM)

* It is suitable for small and medium samples and has good classification effect.
* Use kernel function to deal with nonlinearity.
* Data standardization needs to be done.

### **Data processing and modeling process (R language)**

* 1. Import the data and convert the dependent variable to a factor

# Load the package

library(rpart)

library(randomForest)

library(caret)

library(e1071)

# reading data

data <- read.csv("divorce.csv", header = TRUE, sep = ";", stringsAsFactors = TRUE)

# data preprocessing

# Factorize the category variables

data$Class <- as.factor(data$Class)

* 1. Split the training set (70%) and test set (30%)

# Divide the training and test sets, 70% for training and 30% for testing

trainIndex <- createDataPartition(data$Class, p = 0.7, list = FALSE)

train\_data <- data[trainIndex, ]

test\_data <- data[-trainIndex, ]

* 1. Decision tree, random forest and SVM models are trained separately
     1. decision tree

# Train the decision tree

tree\_model <- rpart(Class ~ ., data = train\_data, method = "class")

tree\_pred <- predict(tree\_model, test\_data, type = "class")

tree\_cm <- confusionMatrix(tree\_pred, test\_data$Class)

* + 1. random forest

# Train the random forest

set.seed(123)

rf\_model <- randomForest(Class ~ ., data = train\_data, ntree = 100)

rf\_pred <- predict(rf\_model, test\_data)

rf\_cm <- confusionMatrix(rf\_pred, test\_data$Class)

* + 1. SVM model

Train support vector machine (SVM)

svm\_model <- svm(Class ~ ., data = train\_data, kernel = "radial")

svm\_pred <- predict(svm\_model, test\_data)

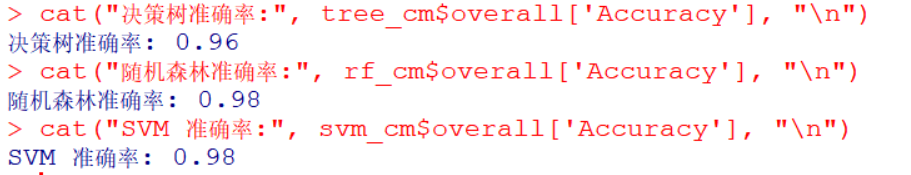
svm\_cm <- confusionMatrix(svm\_pred, test\_data$Class)

* 1. Output accuracy and confusion matrix

Cat ("Decision tree accuracy:", tree\_cm$overall['Accuracy'], "\n")

Cat ("Random forest accuracy:",rf\_cm$overall['Accuracy'], "\n")

Cat ("SVM accuracy:",svm\_cm$overall['Accuracy'], "\n")



confusion matrix:

# If not installed, please install it first

install.packages("ggplot2")

install.packages("reshape2")

library(ggplot2)

library(reshape2)

plot\_conf\_matrix <- function(cm, model\_name) {

cm\_table <- as.data.frame(cm$table)

colnames(cm\_table) <- c("Predicted", "Actual", "Freq")

ggplot(cm\_table, aes(x = Actual, y = Predicted, fill = Freq)) +

geom\_tile(color = "white") +

geom\_text(aes(label = Freq), vjust = 1.5, color = "black", size = 5) +

scale\_fill\_gradient(low = "white", high = "steelblue") +

ggtitle(paste("Confusion Matrix -", model\_name)) +

theme\_minimal()

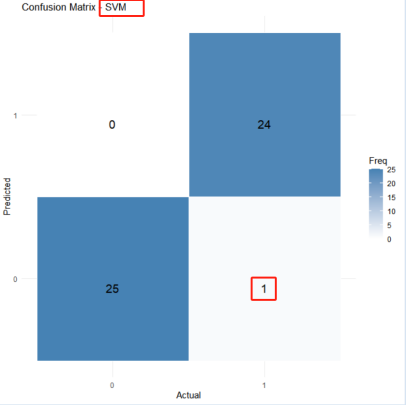
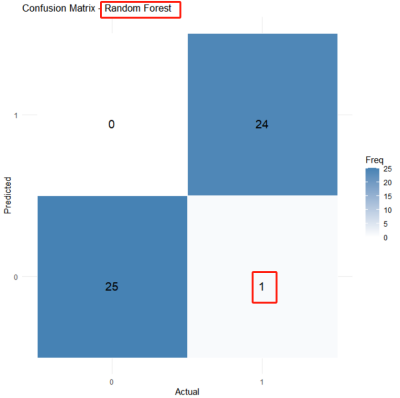
}

# Draw separately

plot\_conf\_matrix(tree\_cm, "Decision Tree")

plot\_conf\_matrix(rf\_cm, "Random Forest")

plot\_conf\_matrix(svm\_cm, "SVM")



From the results, the prediction accuracy of decision tree is 96%, and there are only 2 prediction errors in 50 samples. The prediction accuracy of random forest and SVM is 98%, and there are only 1 prediction error in 50 samples.

* 1. Output feature importance ranking (RF)

# Feature importance extraction

importance\_rf <- importance(rf\_model)

importance\_df <- data.frame(Feature = rownames(importance\_rf), Importance = importance\_rf[,1])

importance\_df <- importance\_df[order(-importance\_df$Importance), ]

# Look at the top 10 important features

head(importance\_df, 10)



### **V. Comparison and evaluation of results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **model** | **Number of samples in the test set** | **Number of errors** | **precision** | **remarks** |
| decision tree | 50 | 2 | 96% | The accuracy is high |
| random forest | 50 | 1 | 98% | More accurate predictions |
| SVM | 50 | 1 | 98% | More accurate predictions |

From the perspective of psychological theory, the top5 importance features are analyzed as follows:

1. Atr17,18: Consistency of common values and goals

Psychological theory: The theory of similar attraction in social psychology suggests that partners with the same values and long-term goals are more likely to maintain a stable relationship. Studies show that differences in values are the second largest predictor of divorce (after communication problems) (citing data from marriage studies).

1. Atr26: Empathy and cognitive understanding

Psychological theory: Empathy is a buffer for marital quality. Knowing your partner's concerns means taking the initiative to pay attention and listen, which is a characteristic of secure attachment; lack of empathy can lead to "emotional alienation," which in turn leads to cold wars or accusations.

3. Atr11: Positive future expectations

Psychological theory: A positive outlook on relationships can improve resilience to conflict. This belief can ease short-term conflicts and avoid "catastrophic thinking" (seeing small problems as a sign of a broken marriage).

4. Atr19: Role allocation consensus

Psychological theory: Social role theory suggests that clear division of roles reduces power struggles. Couples who agree on roles such as "who is responsible for housework/economy/childcare" have 40% less daily friction (citing family research data).

### **6. Summary and improvement direction**

Summary: This project models the prediction of whether marriage is likely to end in divorce using three machine learning algorithms: decision trees, random forests, and SVM. It finds that the accuracy of predictions from the random forest model and SVM is comparable, with an accuracy rate as high as 98%. Given our goal of predicting divorce and providing intervention guidance, we recommend the random forest more, as it can directly output the importance of characteristics affecting marriage, offering better data support for marital counseling and intervention.

But there are also some limitations:

* The small amount of data may affect the model's generalization ability;
* Cultural bias, data may only reflect marriage characteristics of specific regions/cultural backgrounds;
* There is a high degree of correlation between features, and in the future, feature dimensionality reduction (PCA) and other methods can be considered to optimize the model;
* The SVM model can further optimize parameters through grid search.

### **7、 Suggestions and Prospects**

In future research, we can consider:

(1) Using text features and combining NLP analysis of marital conversations to enhance predictive dimensions;

(2) Introducing time series data (such as records of marital interactions);

(3) By combining physiological indicators and increasing body signal data such as heart rate during couple interaction, customers can reflect their emotional state;

(4) Psychological counseling assistance tool, deploy the model as an online questionnaire system, and output a wind direction assessment report;

(5) Curriculum design for marriage education, developing specialized communication training for high importance features (such as Q1/Q2);

(6) Establish a personalized recommendation system for marriage mediation and advice.