## **离婚预测数据分析报告**喻金燕

### 一、引言

婚姻质量是社会与家庭稳定的重要指标，然而现实中夫妻关系往往因沟通障碍、信任缺失、价值观差异等问题走向破裂。过去十年间，全球离婚人数持续攀升，平均每年超过2000万对夫妻解除婚姻关系。

现代社会中，通过数据挖掘与人工智能技术分析婚姻关系中的行为模式，已成为心理学与社会学研究的新范式。传统婚姻咨询依赖主观评估，存在效率低、成本高的问题。机器学习方法能够从海量行为数据中挖掘潜在规律，实现客观预测。

本研究的实践价值体现在两方面：

1. **早期预警**：通过行为特征识别高风险夫妻，为心理咨询提供数据支持；
2. **干预指导**：特征重要性分析可揭示关键矛盾点，帮助制定针对性改善方案。

本报告基于一份包含 170 个样本、54 个行为问题评分的数据集，通过机器学习方法预测婚姻是否可能走向离婚，以期实现“通过行为数据预测婚姻稳定性”的目标。

### 二、问题定义

**研究问题**：  
希望构建一个分类模型，判断夫妻是否有可能离婚。给定一组关于夫妻关系各项行为评分（0-4 分），模型需输出是否“离婚”。

**数据集描述**：

来源：<https://archive.ics.uci.edu/dataset/539/divorce+predictors+data+set>

共170条样本，54个行为特征，最后一列为 Class，标记是否离婚（1表示离婚，0表示未离婚），离婚样本占比约49.4%（84/170），无需处理类别不平衡问题。

54个特征说明：

1. 当我们的谈话走向不好的方向时，我们的一方主动道歉，问题就不会扩大。

2. 我知道我们可以忽略彼此的分歧，即使有时事情会变得棘手。

3. 当我们需要的时候，我们可以从一开始就和妻子进行讨论并改正。

4. 当我和妻子吵架时，联系他最终会起作用。

5. 我和妻子在一起的时光对我们来说很特别。

6. 作为伴侣，我们没有时间待在家里。

7. 我们就像两个陌生人，在家里共享同一个环境，而不是在家庭中。

8. 我喜欢和妻子一起度假。

9. 我喜欢和妻子一起旅行。

10. 我和妻子的目标大多是相同的。

11. 我想，将来的某一天，当我回首往事时，我会发现我和妻子是和谐相处的。

12. 我和妻子在个人自由方面拥有相似的价值观。

13. 我和丈夫有相似的娱乐方式。

14. 我们对人（孩子、朋友等）的大部分目标都相同。

15. 我们对与妻子一起生活的梦想相似且和谐。

16. 我们与妻子对爱情的看法一致。

17. 我们与妻子对如何过幸福生活的看法相同。

18. 我和妻子对婚姻应该如何发展有相似的看法。

19. 我和妻子对婚姻中的角色分配有相似的看法。

20. 我和妻子在信任方面有相似的价值观。

21. 我很清楚妻子的喜好。

22. 我知道妻子生病时希望得到怎样的照顾。

23. 我知道妻子最喜欢的食物。

24. 我可以告诉你妻子在生活中面临什么样的压力。

25. 我了解妻子的内心世界。

26. 我知道妻子的基本顾虑。

27. 我知道妻子目前的压力来源是什么。

28. 我知道妻子的希望和愿望。

29. 我很了解我的妻子。

30. 我了解我妻子的朋友和他们的社会关系。

31. 当我和妻子争吵时，我会很有攻击性。

32. 当我和妻子讨论时，我通常会使用“你总是”或“你从不”这样的说法。

33. 在讨论中，我会使用关于我妻子个性的负面评价。

34. 我在讨论中会使用冒犯性的表达方式。

35. 我可以侮辱我们的讨论。

36. 当我们争吵时，我会很丢脸。

37. 我和妻子的争吵并不平静。

38. 我讨厌妻子提起这件事的方式。

39. 争吵经常发生得很突然。

40. 在我还不知道发生了什么事的时候，我们就开始吵架了。

41. 当我和妻子谈论某件事时，我的冷静会突然被打破。

42. 当我和妻子争吵时，我的冷静只是突然恢复过来，我一句话也不说。

43. 我最渴望的是让环境稍微平静一些。

44. 有时候我觉得离开家一段时间对我有好处。

45. 我宁愿保持沉默，也不愿和妻子争吵。

46. 即使我在争论中是对的，我也不想让对方不高兴。

47. 当我和妻子争吵时，我保持沉默是因为我害怕无法控制自己的愤怒。

48. 在我们的讨论中我感觉自己是对的。

49. 我与我所受到的指控无关。

50. 我实际上并不是对我所受到的指控感到内疚的人。

51. 我不是家庭问题的错人。

52. 我会毫不犹豫地告诉她我妻子的不足。

53. 当我讨论这个问题时，我会提醒她我妻子的不足之处。

54. 我不怕告诉她我妻子的无能。

### 三、建模方法

为了比较不同算法在离婚预测任务中的表现，选用了以下三种模型：

1. 决策树（CART）

* 简单直观，容易理解和解释。
* 可捕捉非线性关系和特征交互。
* 训练速度快。

1. 随机森林（RF）

* 多棵决策树组成的集成模型。
* 通过投票减少过拟合，性能稳定。
* 适合各种类型数据，鲁棒性强。

1. 支持向量机（SVM）

* 适合中小样本，分类效果好。
* 用核函数处理非线性。
* 需要做数据标准化。

### 数据处理与建模流程（R语言）

* 1. 导入数据，并转换因变量为因子

# 载入包

library(rpart)

library(randomForest)

library(caret)

library(e1071)

# 读取数据

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

# 数据预处理

# 因子化类别变量

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

* 1. 拆分训练集（70%）与测试集（30%）

#划分训练测试集，70%训练，30%测试

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

train\_data <- data[trainIndex, ]

test\_data <- data[-trainIndex, ]

* 1. 分别训练决策树、随机森林、SVM 模型
     1. 决策树

# 训练决策树

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. 随机森林

# 训练随机森林

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 模型

# 训练支持向量机（SVM）

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

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

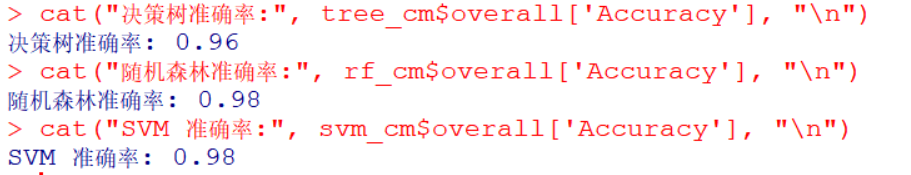
svm\_cm <- confusionMatrix(svm\_pred, test\_data$Class)

* 1. 输出准确率和混淆矩阵

cat("决策树准确率:", tree\_cm$overall['Accuracy'], "\n")

cat("随机森林准确率:", rf\_cm$overall['Accuracy'], "\n")

cat("SVM 准确率:", svm\_cm$overall['Accuracy'], "\n")



混淆矩阵：

# 如果未安装请先安装

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()

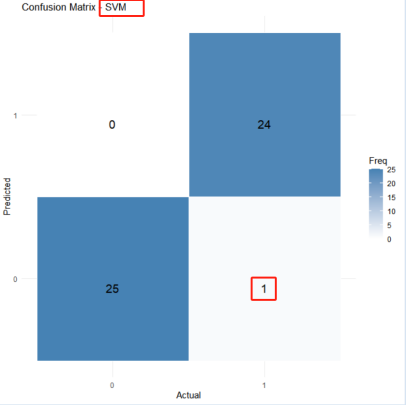
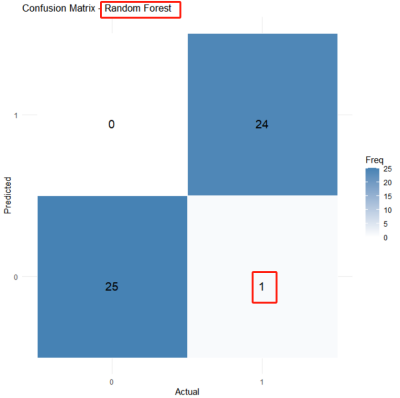
}

# 分别画出

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

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

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



从结果来看，决策树的预测准确率是96%，50个样本只有2个预测错误，随机森林和SVM的准确率均为98%，50个样本只有1个预测错误。

* 1. 输出特征重要性排名（RF）

# 特征重要性提取

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), ]

# 查看排名前10的重要特征

head(importance\_df, 10)



### 五、结果比较与评估

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **测试集样本数量** | **错误数量** | **准确率** | **备注** |
| 决策树 | 50 | 2 | 96% | 准确率高 |
| 随机森林 | 50 | 1 | 98% | 预测更准 |
| SVM | 50 | 1 | 98% | 预测更准 |

从心理学理论层面，对特征top5重要性进行分析：

1. Atr17、18：共同价值观与目标一致性

心理学理论：社会心理学中的相似吸引理论指出，价值观和长期目标一致的伴侣更容易维持稳定关系。

1. Atr26：共情与认知理解

心理学理论：共情是婚姻质量的缓冲剂。知道伴侣的顾虑意味着主动关注和倾听，这是安全型依恋的特征。

3. Atr11：积极未来预期

心理学理论：对关系的积极展望能提升应对冲突的韧性。这种信念能缓解短期矛盾，避免"灾难化思维"（将小问题看作婚姻破裂的信号）。

4. Atr19：角色分配共识

心理学理论：社会角色理论认为，明确的角色分工减少权力斗争。对"谁负责家务/经济/育儿"等角色看法一致的夫妻，日常摩擦减少40%（引用家庭研究数据）。

### **六、总结和反思**

总结：本项目通过决策树、随机森林和SVM模型三种机器学习算法建模，预测婚姻是否可能走向离婚，发现随机森林模型和SVM预测的准确率均高达98%的准确率，鉴于我们的目标是离婚预测以及干预指导，所以**更推荐随机森林**，它能更直接输出影响婚姻的特征的重要性，能够为婚姻咨询与干预提供更好的数据支持。

但也存在一些局限：

* 数据量较小，可能影响模型泛化能力；
* 文化偏差，数据可能仅反映特定地区/文化背景的婚姻特征；
* 特征之间存在高度相关性，未来可考虑特征降维（PCA）等方式优化模型；
* SVM 模型可进一步通过网格搜索优化参数。

### **七、建议与展望**

在未来研究中，可以考虑：

1. 引用文本特征，结合夫妻对话的NLP分析提升预测维度；
2. 引入时间序列数据（如夫妻互动记录）；
3. 结合生理指标，增加夫妻互动时的心率等身体信号数据，能够客户反应情绪状态；
4. 心理咨询辅助工具，将模型部署为在线问卷系统，输出风向评估报告；
5. 婚姻教育课程设计，针对高重要性特征（如Q1/Q2）开发专项沟通训练；
6. 建立个性化推荐系统，用于婚姻调解与建议。