约会数据集分析

数据说明：Speed Dating Data.csv

该数据是 Kaggle 网站提供的芝加哥商学院 Ray Fisman 教授和 Sheena Iyengar 教授 2002 年至 2004 年组织的相亲实验数据。实验的开始，组织者在该校网站上招募相亲者。志愿者需要在网站上注册，经审核后方可参与相亲活动。数据来源为：*https://www.kaggle.com/annavictoria/speed-dating-experiment*

每一位参与者在**约会前**，都需要填写一张调查表，其中包括本人的个人信息（性别、年龄、种族、从事领域和兴趣等）和对理想型的各方面期望打分。在约会后，需要对实际相亲的对象各方面进行打分，并评估出好感度和成功牵手的可能性。该数据集中，每一条观测包含一对相亲成员（两位）的信息，但是是从其中一个人的角度展示的（以下称为“本人”），本人的 ID 是 iid 变量，对方的 ID 是 pid 变量，dec 表示本人（即 iid）做出的决策。

对于**约会后打分**的变量，【被打分】的是 pid，【做出打分这个动作】的是 iid，以 A 和 B 两位参与者为例：

- iid＝A & pid＝B 的数据中，期望打分是 A 在约会前对理想型的描述，约会主观打分是 A 给 B 打的分数；
- iid＝B & pid＝A 的数据中，期望打分是 B 在约会前对理想型的描述，约会主观打分是 B 给 A 打的分数；

完整数据中变量较多，在本次分析任务中将有可能使用到的变量含义及说明如下表所示：

变量类型		变量名	详细说明	取值范围	备注
因变量		dec	决定	定性变量（2 水平）	1 表示是 0 表示否	是否有意愿进一步发展
自变量	参与者编号	iid	本人编号	定性变量	1-552	编号为 118 的观测缺失
		pid	对方编号	定性变量	1-552	编号为 118 的观测缺失
	客观条件	gender	性别	定性变量（2 水平）	1 表示男 0 表示女	-
		age	年龄	单位：岁	18-55	-
		race	种族	定性变量（5 水平）	亚裔、非裔、欧洲裔、拉丁裔、其他	
		field	从事领域	定性变量(多个水平，需要进行进一步筛选、分类)	医学、历史、哲学、商业等	

		一系列表示兴趣的变量: sports（运动）、tvsports（电视运动节目）、exercise（锻炼）、dining（进餐）、museum（博物馆）、art（艺术）、hiking（徒步旅行）、gaming（博彩）、clubbing（逛夜店）、tv（看电视）、theater（戏剧）、movie（电影）、concerts（音乐会）、shopping（购物）、yoga（瑜伽）	单位：分	十分制	只取整数，10 代表最喜欢，1 代表最不喜欢
	对方条件	age_o	对方年龄	单位：岁	18-55	只取整数
		race_o	对方种族	定性变量（5 水平）	亚裔、非裔、欧洲裔、拉丁裔、其他	-
		samerace	是否同一种族	定性变量（2 水平）	1 表示是 0 表示否	-
	约会意愿	go_out	日常出门频率	定性变量（7 水平）	1 表示每周大于两次，2 表示每周两次，3 表示每周一次，4 表示每月两次，5 表示每月一次，6 表示每年若干次，7 表示几乎不	
		date	日常约会频率	定性变量（7 水平）		
		imprelig	对宗教看重程度	单位：分	十分制	只取整数
		imprace	对种族看重程度	单位：分	十分制	只取整数
	*约会前*期望打分	attr1_1	吸引力打分	单位：分	十分制	只取整数
		shar1_1	共同爱好打分	单位：分	十分制	只取整数
		fun1_1	幽默打分	单位：分	十分制	只取整数
		sinc1_1	真诚打分	单位：分	十分制	只取整数
		amb1_1	雄心打分	单位：分	十分制	只取整数
		intel1_1	智力打分	单位：分	十分制	只取整数
	*约会后*主观打分	attr	吸引力打分	单位：分	十分制	只取整数
		shar	共同爱好打分	单位：分	十分制	只取整数
		fun	幽默打分	单位：分	十分制	只取整数
		sinc	真诚打分	单位：分	十分制	只取整数
		amb	雄心打分	单位：分	十分制	只取整数
		intel	智力打分	单位：分	十分制	只取整数
		like	好感打分	单位：分	十分制	只取整数
		prob	评估成功率	单位：分	十分制	只取整数

分析任务：

1. 读入数据，并选出上述表格所示的变量，并确定一个你感兴趣的选题方向；例如：什么因素与两人进一步发展有关？
2. 完成与你的选题相关的 2-3 个描述分析；形式：参考课上讲的描述分析内容；

3. 针对选题目的进行建模，并解读模型的结果；需要包含模型：决策树（需解读模型结果）、Boosting 模型、随机森林、SVM 模型的比较；
提示：可使用 R package: tree, gbm, randomForest, e1071::svm()实现相关函数
4. 结合任务 3-4 的结果，得出你的结论。

描述分析参考资料：

- 《狗熊会 | 丑图百讲》：你画过的丑图，这里全都有
- https://www.r-graph-gallery.com/index.html r-gallery 提供了非常丰富的作图模板 如果你不知道怎么展示自己的图形 仿照优秀作品是最好的学习方式 里面提供了代码 作图案例等
- https://python-graph-gallery.com/：作用同上

建模部分参考资料：

- 《狗熊会 | Python 数据科学实践》
- 《狗熊会 | R 语言内容实战》
- 阿里天池 AI 平台：数据挖掘 > 07-01----07-06 数据分析项目实践