R作业7——泰坦尼克号生存预测互评

lucifercook 2017.8.28

本次的任务是使用Kaggle上的泰坦尼克号生存数据，使用机器学习的方法建立预测模型，之后对测试集的数据进行预测

第一步是进行数据探测，初步查看各变量和是否幸存的关系，下一步是对缺失值进行处理，因为补充的结果要用于预测，所以如何补充缺失值非常关键，本例中可以根据其他特征值对年龄进行预测。

下一步是将数据转化成适合模型的类型，像随机森林就不能处理charctor型的变量，需要转成factor或者int型。

然后就是选择合适的模型，不同模型在不同的问题中会有不同的表现，因此若要得到比较好的结果，可以多用几个模型然后做比较，选择其中最好的用来正式预测，本次大家普遍使用了随机森林的模型，只有Han Wang同学使用了SVM和神经网络模型，在本例中，SVM有较佳的表现。对同一个模型来说，仍然可以通过调整参数来进一步提升准确率，因为涉及到较多更深入的知识，需要继续研究