Task4:

选用XGBoost算法比较好

LR:

缺点：1）容易欠拟合，分类精度不高。2）数据特征有缺失或者特征空间很大时表现效果并不好

Random Forest：

随机森林已经被证明在某些噪音较大的分类或回归问题上会过拟

对于有不同取值的属性的数据，取值划分较多的属性会对随机森林产生更大的影响，所以随机森林在这种数据上产出的属性权值是不可信的。

XGBoost：

xgBoosting借鉴RF的做法，支持列抽样，这样不仅能防止过拟合，还能降低计算；xgBoosting的代价函数引入正则化项，控制了模型的复杂度，正则化项包含全部叶子节点的个数，每个叶子节点输出的score的L2模的平方和。从贝叶斯方差角度考虑，正则项降低了模型的方差，防止模型过拟合

对于这组缺失率较高，特征容量大的数据来说，XGBoosting无疑表现更加出色

Task5：

需求分析：听取相关业务部门的需求，了解业务所需要的相关数据。以感冒为例，我们需要搜集患者的临床表现比如：是否发热，是否咳嗽，是否流鼻涕，时间，睡眠时间等特征，以及医生的诊断结果等进行工程前的准备。

采集数据：需要相关部门提供数据或是利用爬虫寻找相应的数据

建立模型：需要设计者根据业务设计适合的数学模型来分析数据

训练，测试模型：训练类和测试类来训练模型，调整参数，防止过度拟合和调整幅度过大或过小导致模型预测不准。

上线系统，听取反馈：将系统应用于实际生产中，收集相关部门反馈，寻求改进方案，完善系统设计。

