工作记录6之R中的变量筛选

2016-09-06 16:06:18

建立违约预测模型的过程中，变量的筛选尤为重要 粗筛：随机森林 细筛：WOE转化+决策树模型 随机森林模型 randomForest包或party包 library(party)　#与randomForest包不同之处在于，party可以处理缺失值　set.seed(1)　crf <- cforest(y~.,control = cforest_unbiased(mtry = 2, ntree = 50), data=step2_1)　varimpt <- data.frame(varimp(crf))　party包中的随机森林建模函数为cforest函数，mtry代表在每一棵树的每个节点处随机抽取mtry 个特征，通过计算每个特征蕴含的信息量，特征中选择一个最具有分类能力的特征进行节点分裂，varimp代表重要性函数　woe转化 library(devtools)　install_github("riv","tomasgreif")　#install_github("tomasgreif/riv") library(woe) IV <- iv.mult(step2_2,"y",TRUE)　#原理是以Y作为被解释变量，其他作为解释变量，建立决策树模型 iv.plot.summary(IV) summary(step2_3) 详见博主http://blog.csdn.net/sinat_26917383/article/details/51728515 不过大部分文章。。其实看着意义不太大。。应该博主自己也所知有限 随机森林的官方说明： https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm