天善智能
TIANSHAN SOFT
专注于商业智能BI和大数据的垂直社区平台

分类预测及模型评估

谢佳标（Daniel.xie）

分类及模型评估

- 分类算法是基于类标号的训练集数据建立分类模型并使用其对新观测值（测试数据集）进行分类的算法，属于有监督学习方法。

- 对于分类来说，需要把精力花费在学习问题找到合适的分类器。这时候，考虑到不同算法间的各种差异是很有帮助的。例如，在分类问题中，决策树因在建模过程中有明确的规则输出使得模型通俗易懂，而黑箱操作的神经网络得到的模型则很难解释。 如果你要设计一个欺诈用户甄别模型，上述的模型特点就是很重要的选择模型依据，因为需要发现欺诈用户的异常行为模式，即使神经网络算法能更好地甄别欺诈用户，但是如果不能解释清楚这些预测背后的模式，那么再好的预测也是没有用的。

常用分类算法

- KNN近邻分类
- 朴素贝叶斯分类
- 决策树模型
- 集成学习与随机森林
- 人工神经网络与支持向量机

模型评估

- 为了确保模型能够对未知对象进行正确预测，需要对模型性能进行评估，避免模型出现可能存在的过拟合问题。我们可以利用包括caret、rminer和rocr这些算法包来评估模型性能以防止过拟合问题。进一步，对模型性能进行评估还有助于我们得到更优、健壮性更好的模型，以便准确地预测未知数据。在数据挖掘过程中，每个模型和算法具有不同的适用场景，但最终评价的结果一定是以业务价值为导向，而非模型自身的评价指标，所以，在选取模型时，将更多的将采取一些机器学习算法。

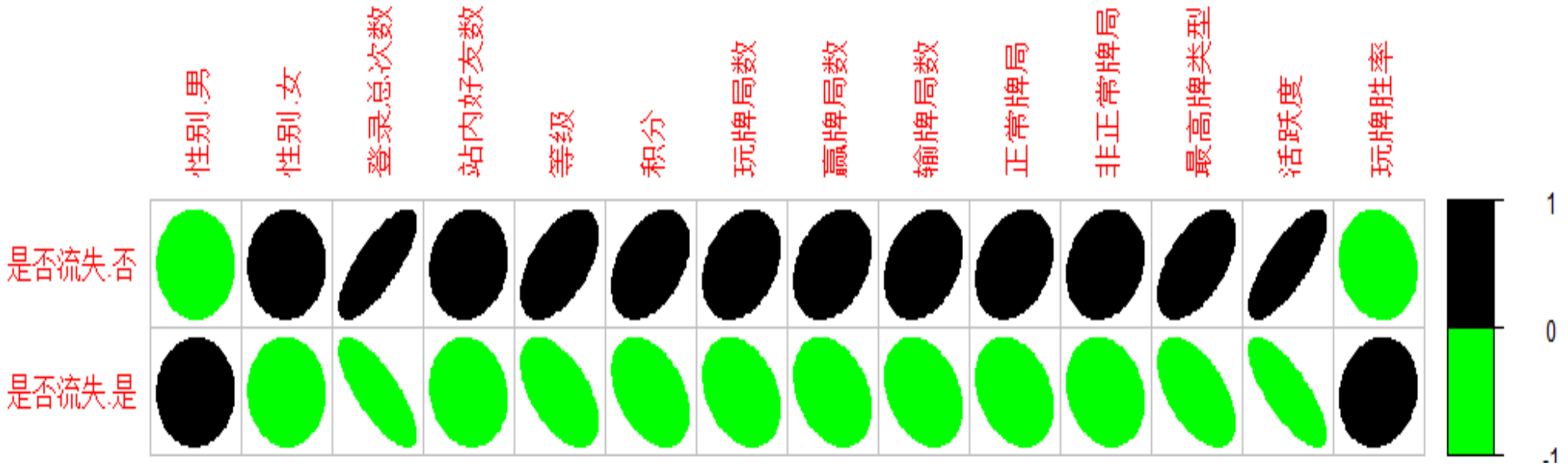
业务案例：活跃用户流失预测

影响活跃用户流失的普遍判断有：在线活跃、用户账号属性（性别、好友数、等级、积分等）和玩牌情况（玩牌局数、赢牌局数、输牌局数、最高牌型等）。

- 用户 id
- 是否流失
- 性别
- 登录总次数
- 站内好友数
- 等级
- 积分
- 玩牌局数
- 赢牌局数
- 输牌局数
- 正常牌局
- 非正常牌局
- 最高牌类型

业务案例：活跃用户流失预测

- 数据探索：研究是否流失与其他变量间的关系？



业务案例：活跃用户流失预测

- 在建模之前，也可以利用10折交叉验证的方法来对参数进行最优选择，此处我们以决策树、随机森林和人工神经网络算法的参数选择为例进行演示。

```
> # 利用10折交叉验证来选择最优参数
> control <- trainControl(method="repeatedcv",number=10,repeats=3)
> rpart.model <- train(是否流失~.,data=w,method="rpart",
+                      trControl=control)
> rf.model <- train(是否流失~.,data=w,method="rf",
+                   trControl=control)
> nnet.model <- train(是否流失~.,data=w,method="nnet",
+                      trControl=control)
```

业务案例：活跃用户流失预测

- 最后，根据找出的最优参数利用算法建立不同的分类器，并对训练集和测试集数据进行预测，通过混淆矩阵查看错误率，从而找出最优模型对其他样本进行活跃流失预测。