|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | |  | |  | |  | |
| Mojing Project 研发文档 |
| Davy Zhu  SH DreamStudio April 4, 2016 |

目录

Mojing Project 研发文档 1

1

数据概况 1

数据预处理 2

建模 2

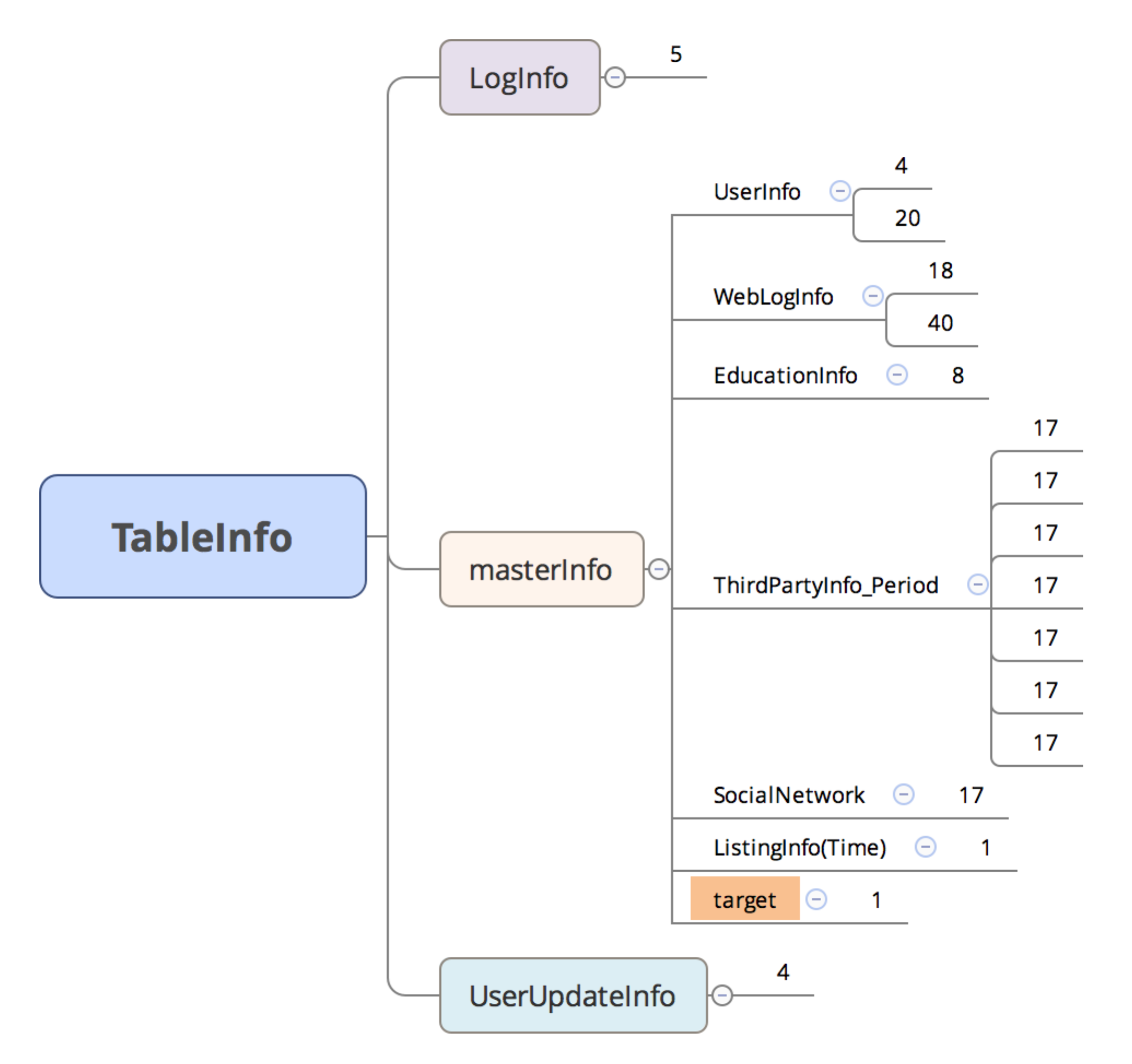
模型评估 3

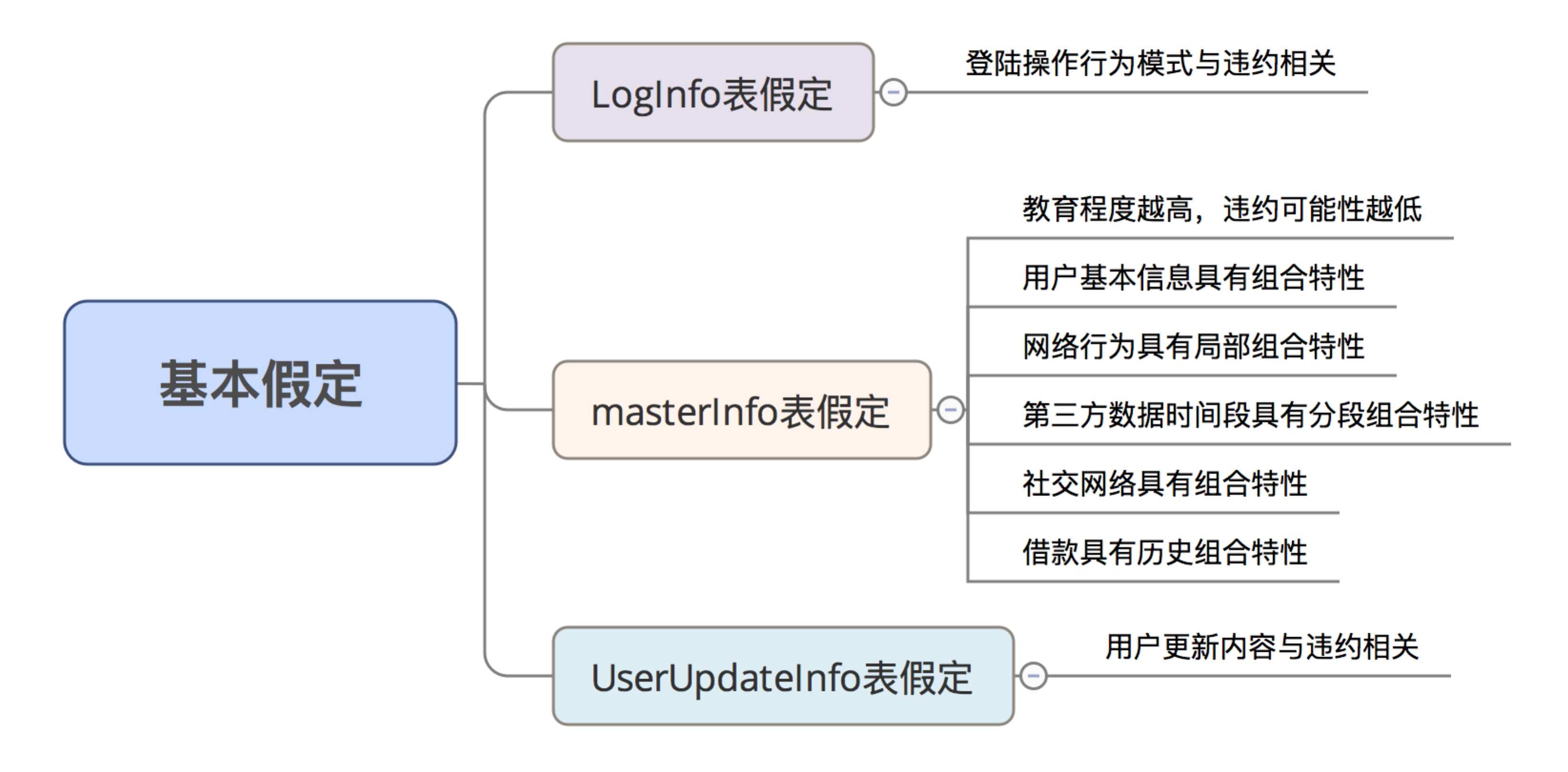
附录 4

# Mojing Project 研发文档

## 

### 数据概况





### 数据预处理

对训练集MasterInfo进行基本分析，获得空变量信息如下表所示：

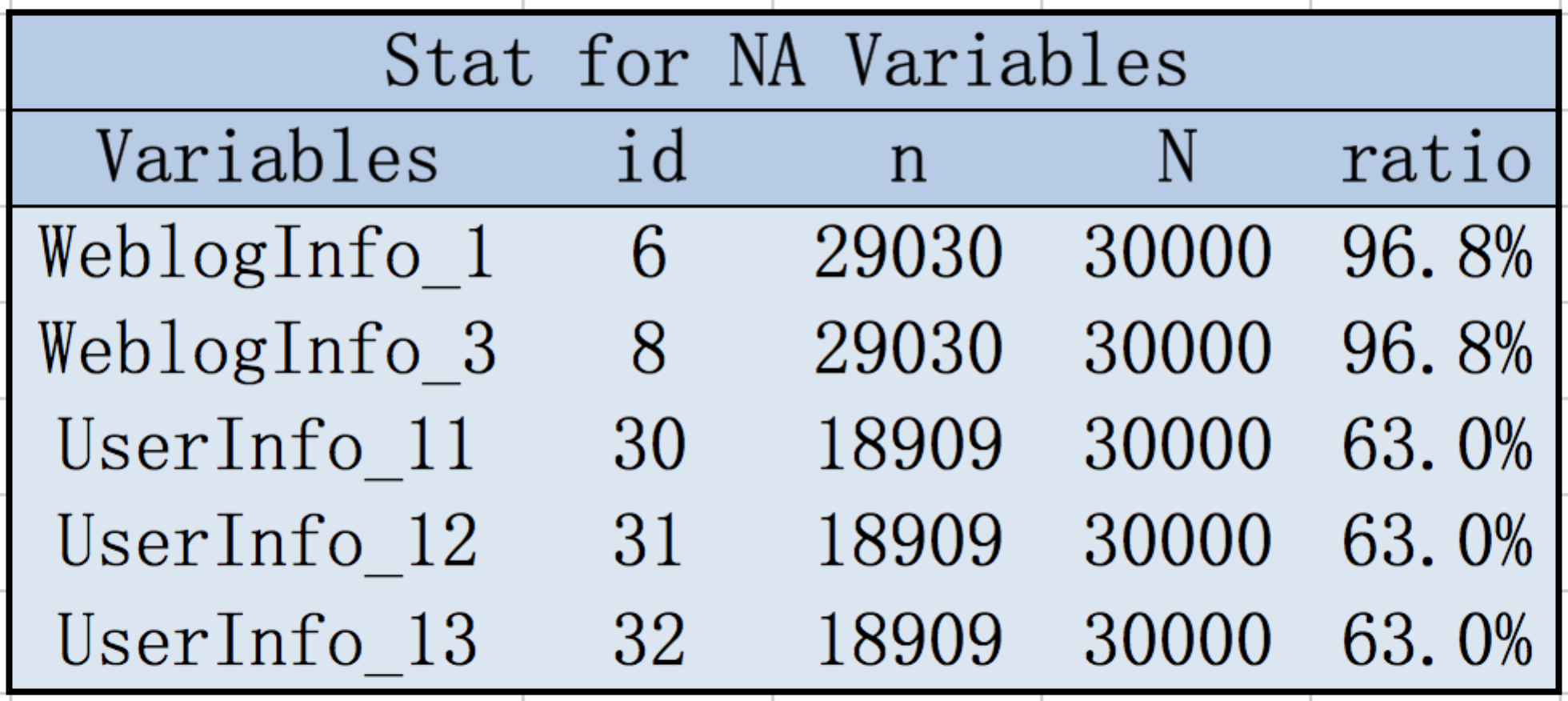
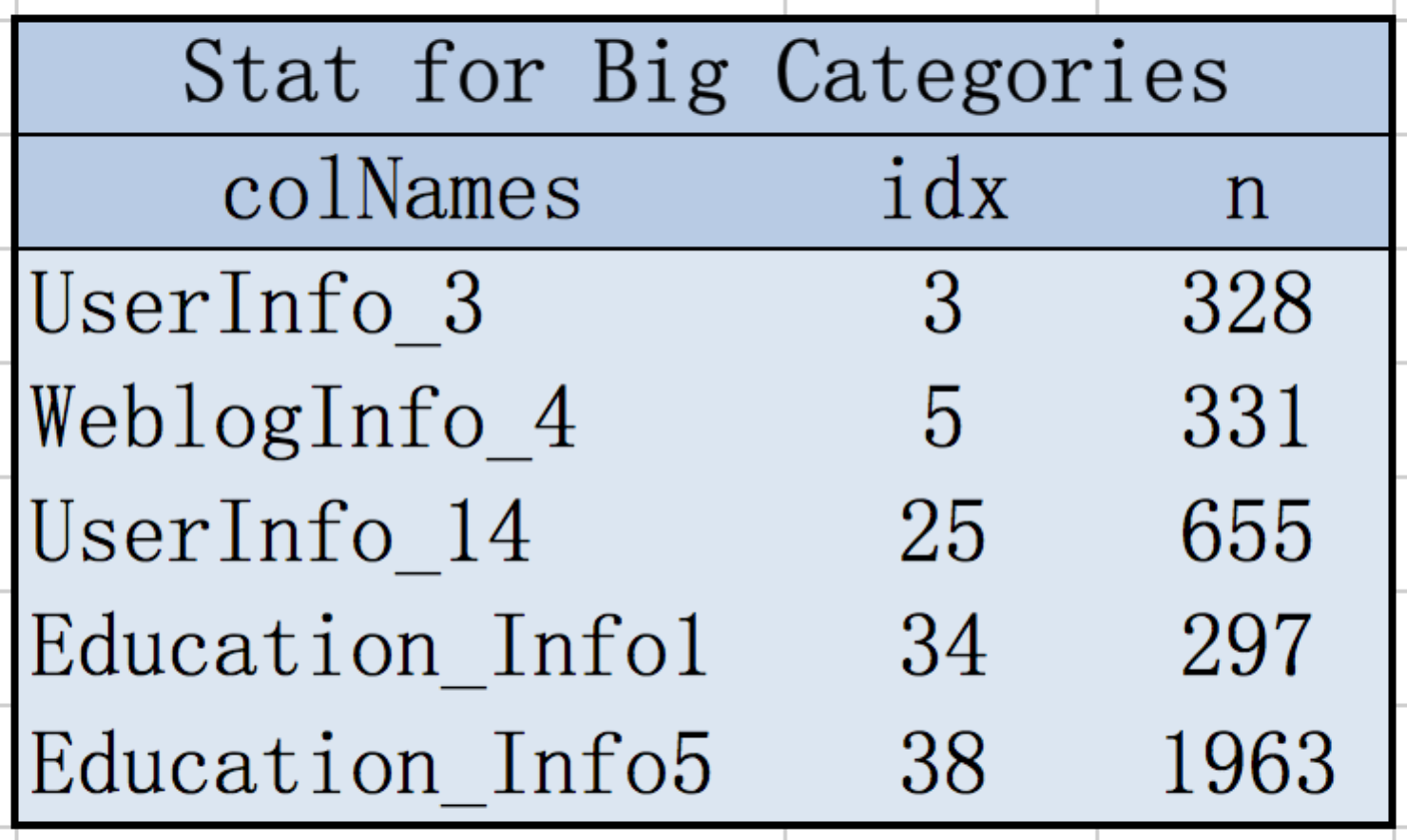
 

表1 表2

又对各category字段进行统计，可以获得多类别型字段(50个类别以上)，如表2所示。剔除掉所有空值大于50%的列，把多类别字段同其它字段分开来构建模型。

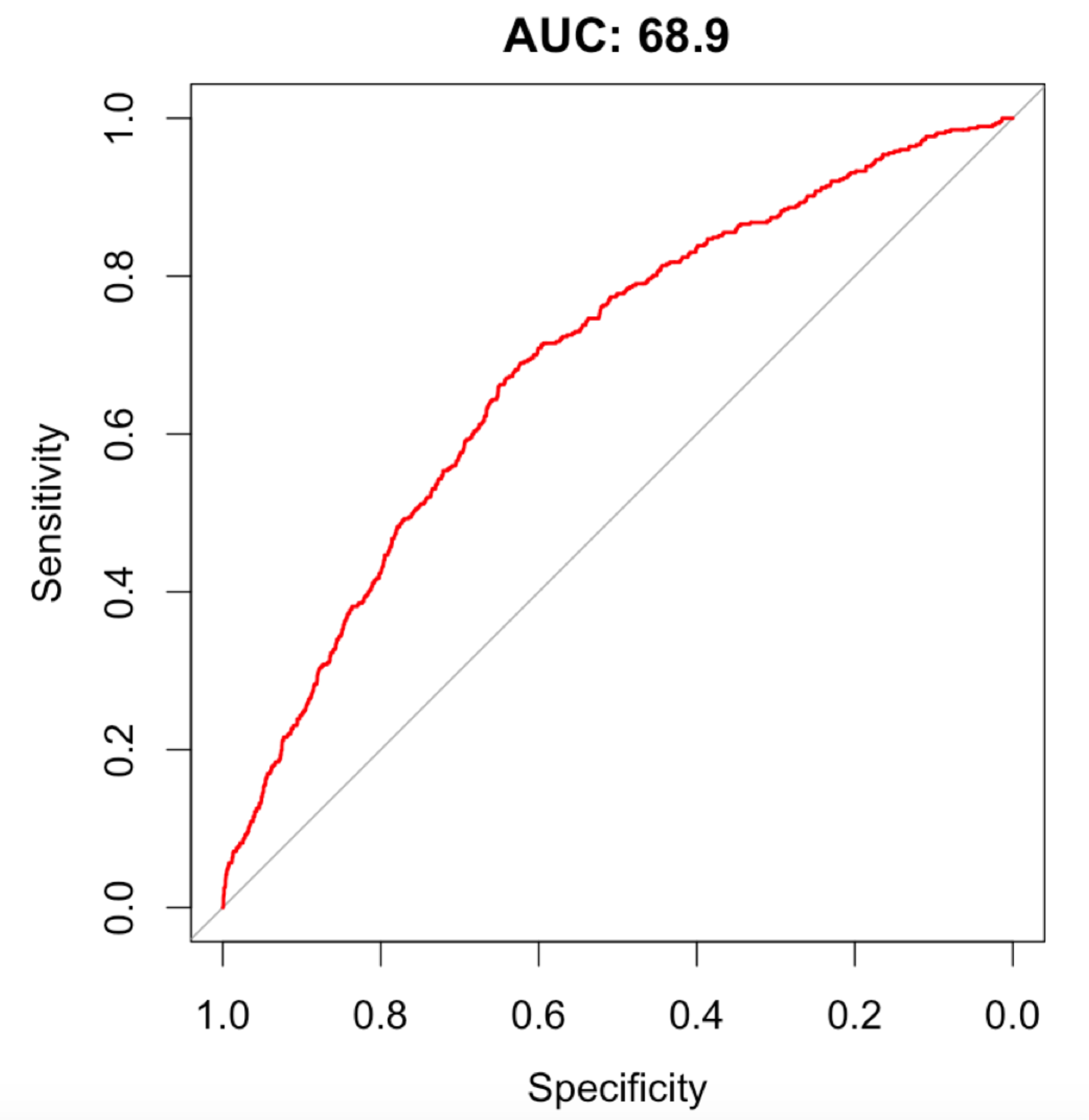
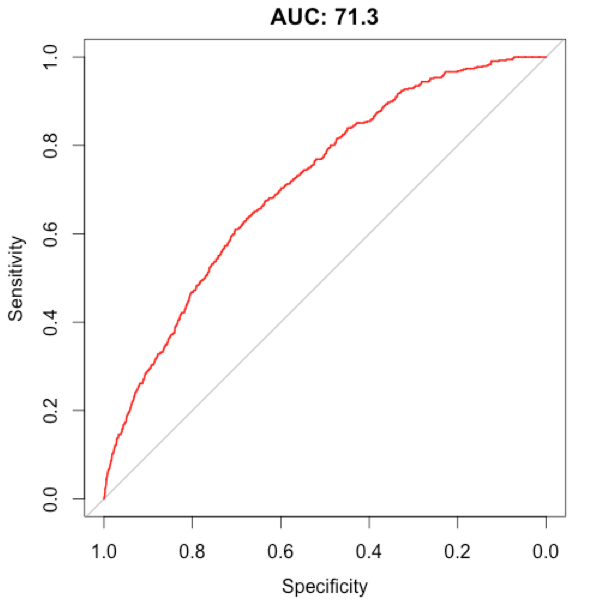
### 建模

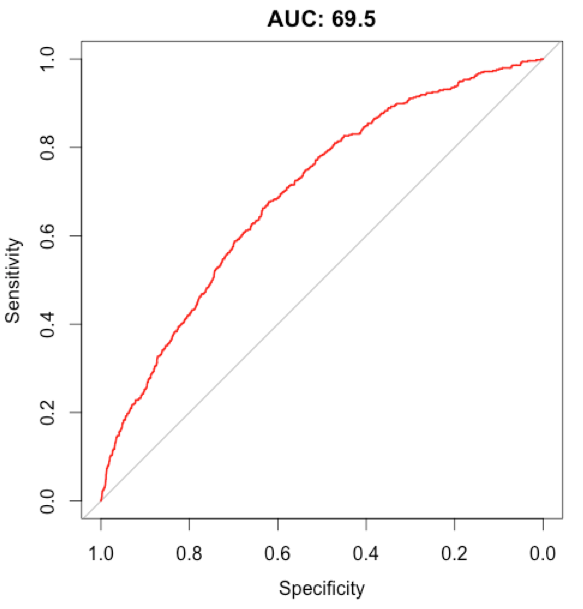
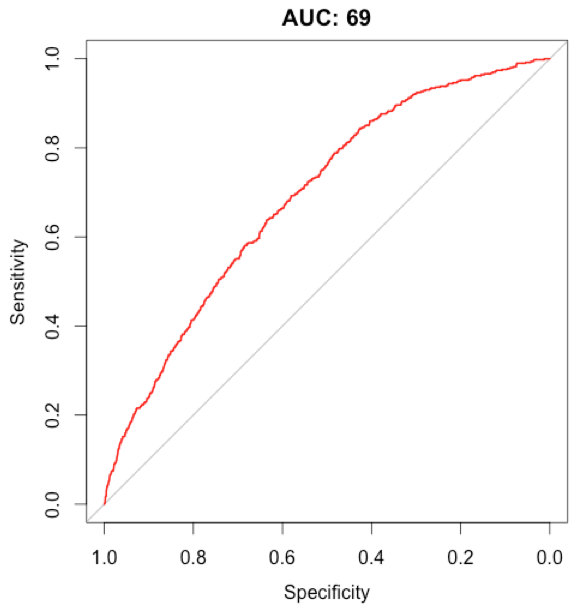
经调研和考虑，选用随机森林(Random Forest)方法对MasterInfo进行建模。对列数固定数值进行采样，对行进行有放回采样30000次，训练500个完全决策树模型进行集成。

### 模型评估

因测试集未给出label，因此只能对训练集进行模型好坏的初步估计，采用AUC来衡量模型好坏。

* 第一版采用masterInfo表中剔除大类别和多空值字段进行训练，约1.7小时，AUC为0.5，其中recall＝0，说明严重underfitting。
* 第二版根据第一版random forest结果，获得对target为1的重要字段，并进一步剔除不重要字段进行训练，约34分钟，AUC为1，说明overfitting。

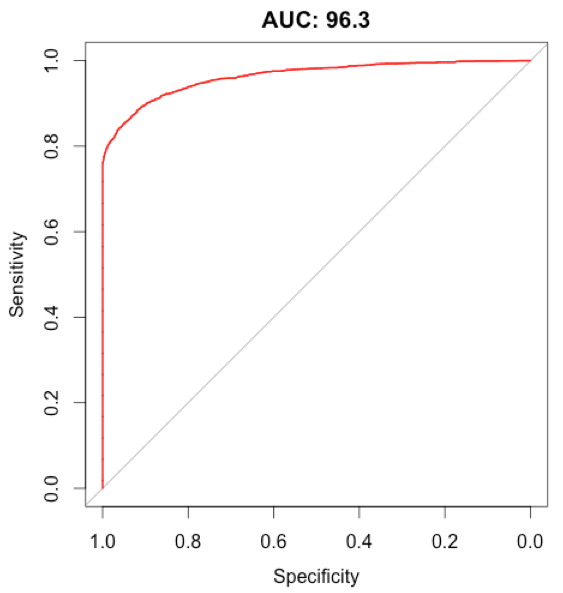
 

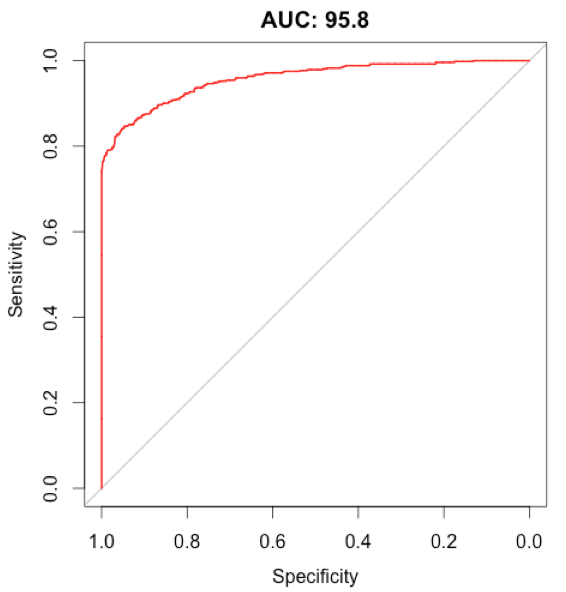
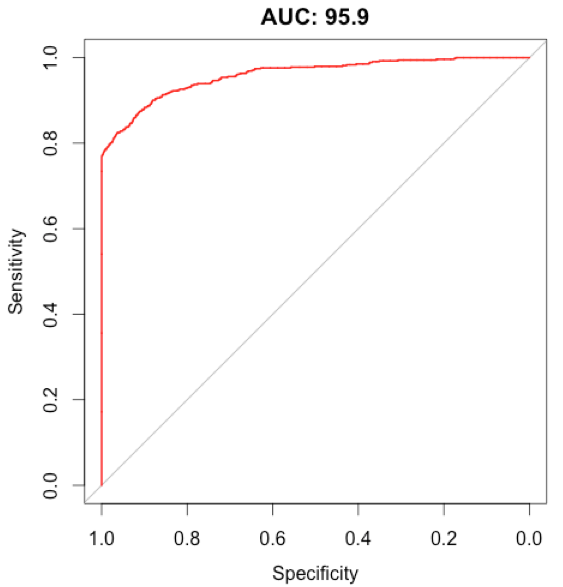
* 第三版把training data分割成两部分(训练集和数据集为3:1)，进行多次验证计算，平均每次约25分钟，AUC分别为68.9，71.3，69.5，69。因此平均AUC为69.6

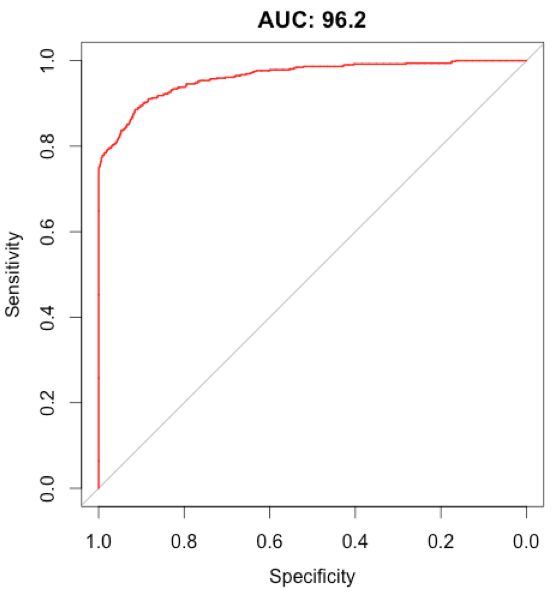
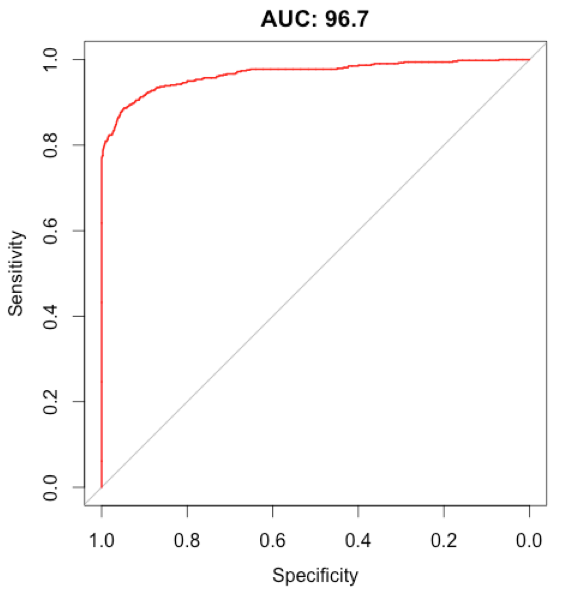
### 附录

因测试集未给出label，因此只能对训练集进行模型好坏的初步估计，采用AUC来衡量模型好坏。

1. 当读者打开折页册时，这将是他们最先看到的文字，此处是简短扼要地摘要您所提供之产品或服务的好位置。
2. 请务必让简介的范围保持精简，如此您才能在折页册的有限篇幅中充分涵盖您所提出的概念。
3. 有几种方式可以组织折页册的内容。您可以选择在每一栏放置不同的要点或主题，如品质及数值。请记住这些要点必须要与折页册首页的介绍文字相结合。
4. 

此区域可用来提供您希望读者采取之下一步骤的清楚指示。可能是您要他们拨打的电话号码、希望他们造访的网站，或是希望他们填写的资讯。无论如何，这份资讯应该要清楚、简短并能提供读者继续往下阅读的动机。­­

参考文献

姓，（日期）*书籍*，页码。

姓，（日期）*书籍*，页码。

姓，（日期）*书籍*，页码。