**大数据挖掘与机器学习**

**课程设计**

**题 目** 银行精准营销客户存款预测

**学 院** 商学院

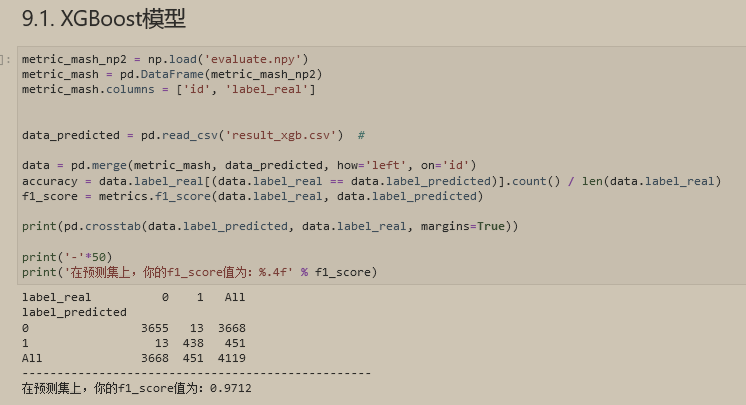
**专 业** 大数据管理与应用

**班 级** 数 据2 0 0 1

**学 生** 陈星宇

**指导教师** 蔡永明

**二〇二二年七月二日**



目录

[1. 原始数据与理解 1](#_Toc107395883)

[1.1 原始数据及含义 1](#_Toc107395884)

[1.2 原始数据理解 2](#_Toc107395885)

[1.3 缺失值(Unknown值)处理 3](#_Toc107395886)

[2. 关联性可视化分析和数据处理 4](#_Toc107395887)

[2.1 Object数据与label 4](#_Toc107395888)

[2.2 数值型数据与label 5](#_Toc107395889)

[2.3 列删除 7](#_Toc107395890)

[3. 异常值可视化 8](#_Toc107395891)

[4. 离散型数据编码 10](#_Toc107395892)

[5. 数据建模 12](#_Toc107395893)

[5.1 模型参数 12](#_Toc107395894)

[5.2 网格搜素算法调参和手动调参 13](#_Toc107395895)

[6. 最终结果 16](#_Toc107395896)

1. 原始数据与理解

## 1.1 原始数据及含义

数据集

1. bank: 训练集，用于训练模型，共41188条数据
2. bank\_prediction: 预测集，用于预测y值，共4119条数据

数值型字段

|  |  |
| --- | --- |
| age | 年龄 |
| duration | 职业 |
| campaign | 本次营销活动中与客户联系的次数 |
| pdays | 距离上次联系客户的天数 |
| previous | 先前活动中与客户联系的次数 |
| emp.var.rate | 季度就业变化率 |
| cons.price.idx | 月度消费者价格指数 |
| euribor3m | 三月期利率 |
| nr.employed | 员工人数 |
| id | 编号 |

object型字段

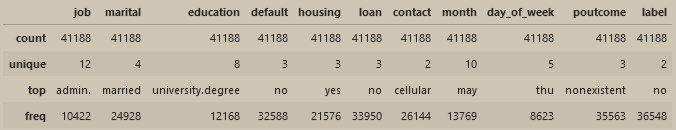
|  |  |
| --- | --- |
| job | 工作 |
| marital | 婚姻状况 |
| education | 教育水平 |
| default | 信用卡状况 |
| housing | 房贷状况 |
| loan | 个人贷款状况 |
| contact | 通话媒介 |
| month | 上次联系时所处的月份 |
| day\_of\_week | 上次联系时为周几 |
| poutcome | 先前活动的结果 |
| label | 是否会购买银行的定期存款业务 |

## 原始数据理解

数值型数据基本情况

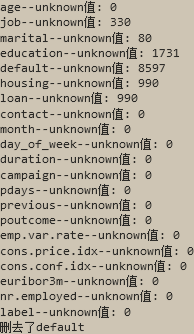


Object型数据基本情况



## 缺失值(Unknown值)处理

bank数据集不存在空缺值，但是存在“Unknown”值，这里对缺失值大于20%的列进行删除操作，这里default列存在缺失值条数为8597，占总条目的约21%，进行删除。

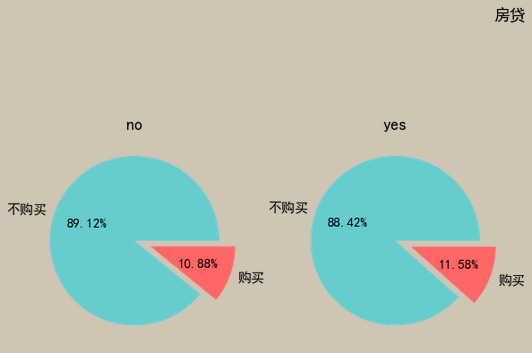
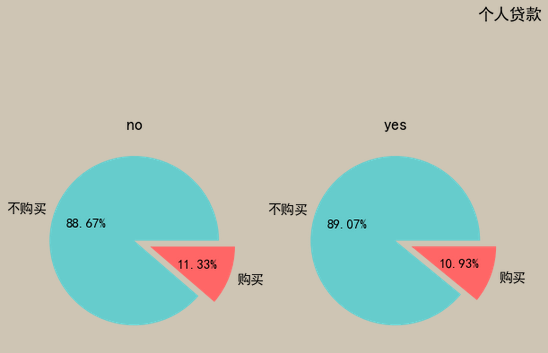
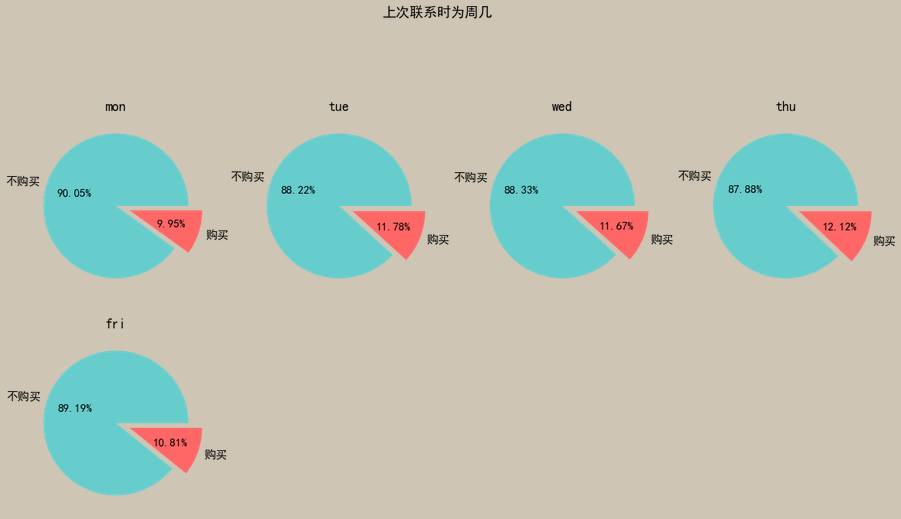


1. 关联性可视化分析和数据处理

## 2.1 Object数据与label

以饼状图的形式分析Object型数据job/marital/education/default(已删除)/housing/loan/contact/month/day\_of\_week/poutcome与label的影响情况。

经过可视化发现：

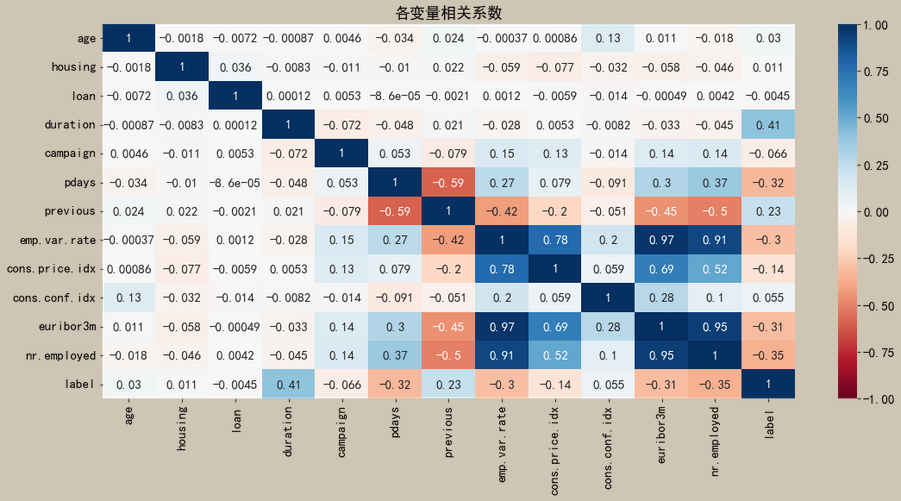
1. 购买定期存款的人数有11.58%有房贷，未购买定期存款的同样有10.88%有房贷，即房贷(housing)可能与label没有直接联系。
2. 同理可得出，个人贷款(loan)、上周联系时间(day\_of\_week)也与label没有直接联系。

## 数值型数据与label

使用热力图方法来描绘数值型数据之间的联系，着重注意较大正值（正相关）和较小负值（负相关），需要注意的是值较小并不代表两个值之间无关，两者可能存在着隐性关系。

通过热力图发现：

1. emp.var.rate与euribor3m相关性为 0.97；emp.var.tate与nr.employed相关性为 0.91；nr.employed与enribor3m相关性为0.95，三者之间彼此都有很强的关系，并且对label影响都在-0.3~-0.4之间，可以删除一列或两列以避免模型重复学习导致因多重共线性造成的过拟合。
2. 与label弱相关的有：loan(-0.0045)、campaign(-0.066)、cons.conf.idx(0.055)，可以尝试删除。



## 列删除

根据以上分析和后期模型校验, 这里删除的列有：

1. 删除emp.var.rate和nr.employed，原因是emp.var.rate和euribor3m和nr.employed三者高度线性相关
2. 删除campaign，通过热力图分析出与label关系性不高

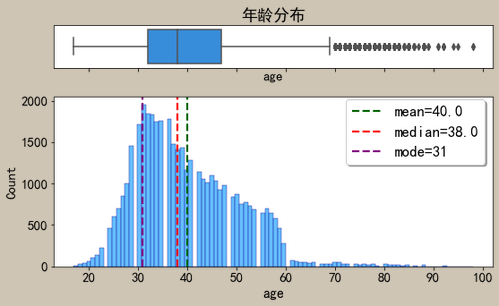
饼图可视化函数：

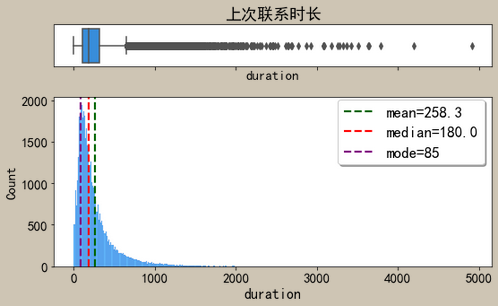


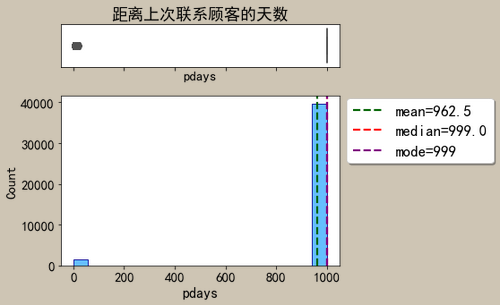
1. 异常值可视化

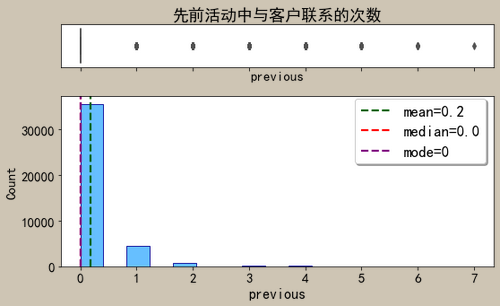
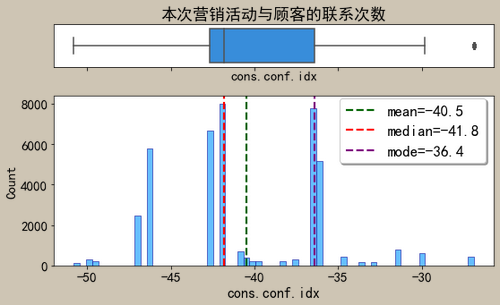
这里主要通过箱线图和分布图可视化异常值情况，如果按照1.5IQR标准确实出现了不少异常值，但是没有出现不合理的数值，故不进行处理。

这里仅展示出使用1.5IQR校验出异常值的列。







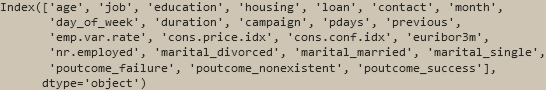


可视化函数：



1. 离散型数据编码

使用序数编码和独热码进行编码，对于education、month、day\_of\_week、job进行序数编码，对marital、poutcome使用独热编码。



独热码编码（marital）



序数编码（education）



1. 数据建模

这里使用XGBoost和CatBoost这两个模型，最终选定XGBoost作为最终模型使用，这里尝试调节了以下几个参数。

## 5.1 模型参数

|  |  |
| --- | --- |
| learning\_rate | 学习率,通过减少每一步的权重，可以提高模型的鲁棒性 |
| min\_child\_weight | 决定最小叶子节点样本权重和。这个参数用于避免过拟合。当它的值较大时，可以避免模型学习到局部的特殊样本。但是如果这个值过高，会导致欠拟合。这个参数需要使用CV来调整。 |
| max\_depth | 树的深度，值越大，越容易过拟合；值越小，越容易欠拟合。默认值为6，典型值3-10 |
| gmma | 在节点分裂时，只有分裂后损失函数的值下降了，才会分裂这个节点。Gamma指定了节点分裂所需的最小损失函数下降值。这个参数的值越大，算法越保守。这个参数的值和损失函数息息相关，所以是需要调整的 |
| subsample | 训练每棵树时，使用的数据占全部训练集的比例导致过拟合。默认值为1 |
| colsample\_bytree | 训练每棵树时，使用的特征占全部特征的比例，防止过拟合 |

## 5.2 网格搜素算法调参和手动调参

以测试集的f1\_score作为评判标准，learnning\_rate=0.07，n\_estimators=100， min\_child\_weight=1，max\_depth=6，gamma=0，subsample=0.7，colsample\_bytree=1时在测试集有最高f1\_score值，但数值仍然较低只有0.60，在测试集表现也不佳。

其它数值不变，当把learning\_rate调节至1.15时在测试集的f1\_score有最高值0.9712，在测试集的f1\_score值约为0.55。通过调节可以把测试集的f1\_score值调节为0.68，但是预测值f1\_score反而下降。

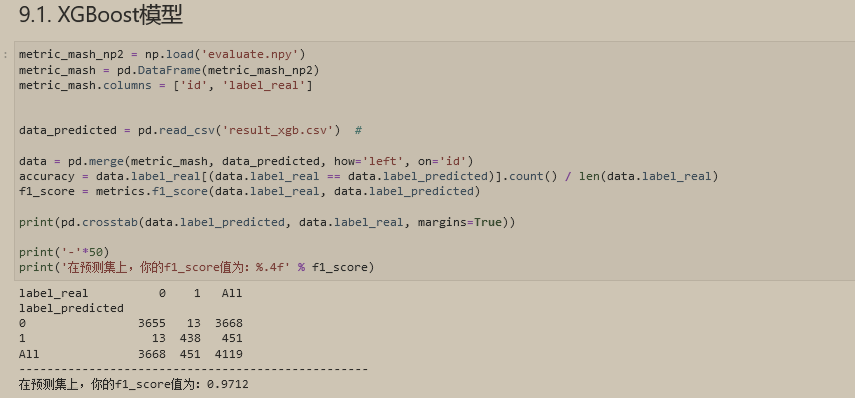
最终使用的参数为learning\_rate=1.15,其它参数使用默认值。

网格调参函数：





1. 最终结果



参考网站：

<https://blog.csdn.net/weixin_39653948/article/details/113823059>

<https://blog.csdn.net/a284365/article/details/116165723>

<https://zhuanlan.zhihu.com/p/95304498>

<https://blog.csdn.net/qq_46000115/article/details/123194498>