|  |
| --- |
| 课程成绩： |
| 任课教师：郑子彬 |



**《数据挖掘项目实验报告》**

|  |  |
| --- | --- |
| 题目： | 基于移动网络通讯行为的风险用户识别 |
| 完成时间： | 2018.6 |
| 学院，专业： | 数据科学与计算机学院  计算机科学与技术专业 |
| 用户名： | Dm15336195 |
| 成绩： | 0.762272 |
| 排名： | Rank = 128st / 211 |
| 组员&贡献率： | 吴立渝 |

1. 实验平台:

语言:python –jupyter notebook

模块:pandas、numpy、lightgbm、sklearn、seaborn、matplotlib

1. 项目目标:

以模拟的语音通话、短信收发、网站及App访问记录等移动网络使用行为为基础，通过数据提取技术和机器学习算法，构建识别风险用户的预测模型，判别用户属于风险用户的可能性。从而为各行业提供风控保障，助力新时代大数字生态的健康有序发展。

1. 项目内容:

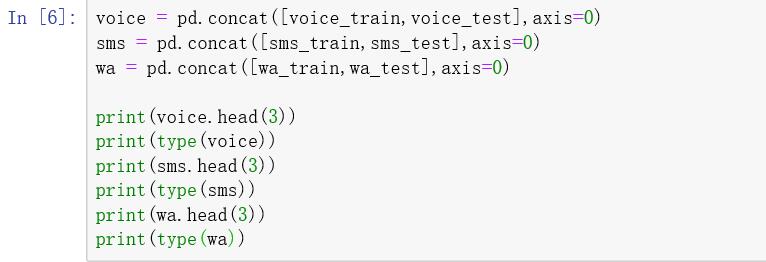
关于本次项目的数据详情:

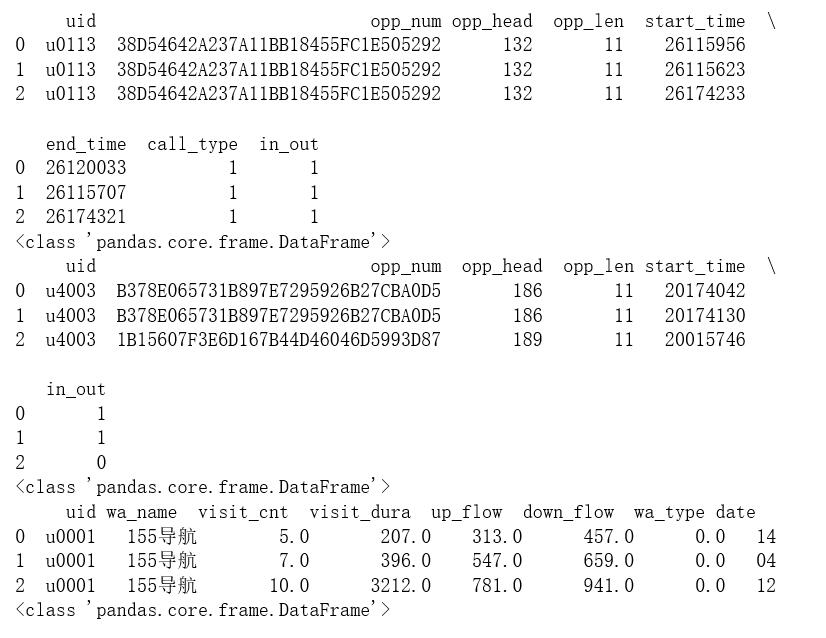


* 数据预处理:

首先，分别读入本次实验给出的数据，为了提高效率，我们使用pandas库中的concat函数将训练集和测试集合并同时提取特征。

在合并后，分别观察给出的特征以及合并后是否为dataframe对象，方便之后的数据处理。(具体代码会在附件中)

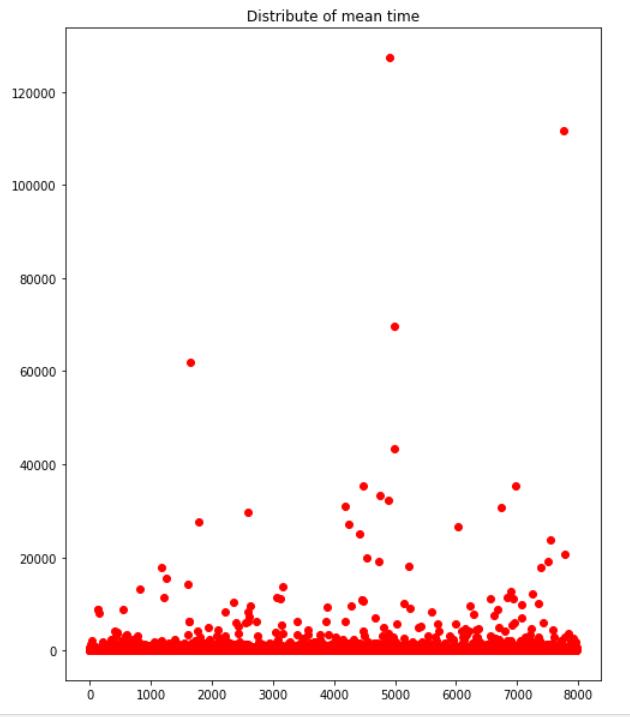
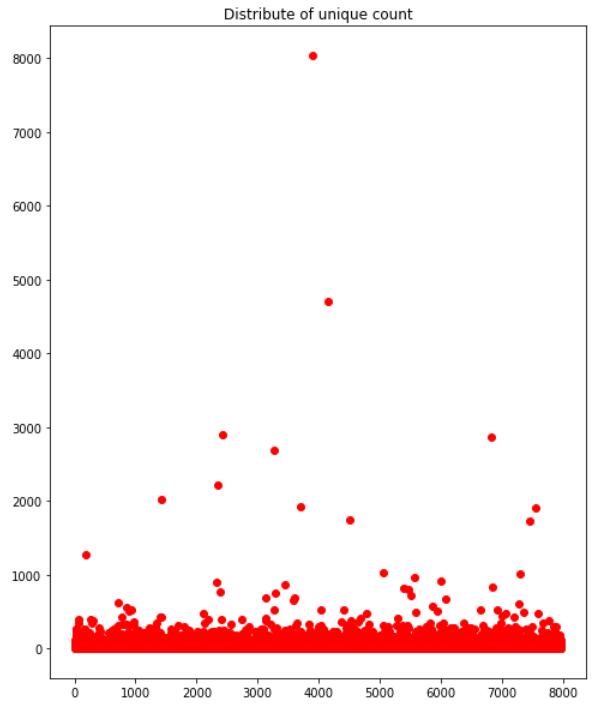




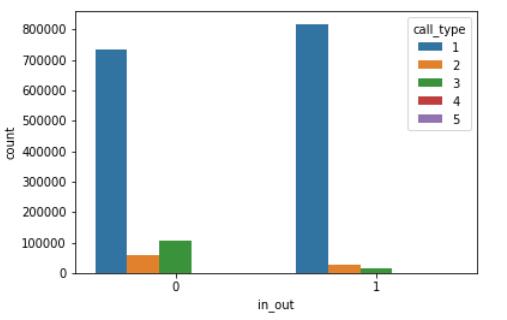
* voice:

关于voice模块，从中提取7个特征，分别为voice\_opp\_count、voice\_opp\_unique\_count、voice\_opp\_head\_ unique\_count、voice\_opp\_len、voice\_mean\_time、voice\_call\_type、voice\_in\_out。

其中，我们将voice\_opp\_unique\_count、voice\_mean\_time的分布进行可视化，查看是否有明显的异常点。可以看到有少数的数据分布较突出，所以在模型训练时，参数需要稍作调正避免因异常点造成分类问题，目前python中sklearn或本次使用的lightgbm都有考虑对异常点的包容



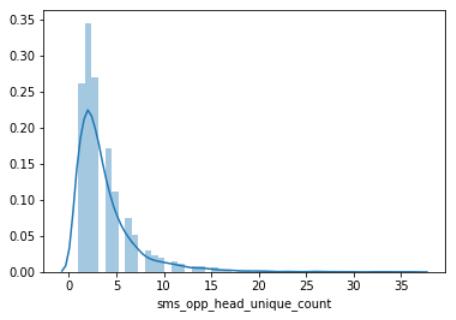
下突为根据in\_out分类进行call\_type的统计，可以看出不管是打进或打出，本地电话占大多数。



* sms:

关于sms模块，从中提取5个特征，分别为sms\_opp\_num\_unique\_count、sms\_opp\_num\_count、sms\_opp\_head\_unique\_count、sms\_opp\_len、sms\_in\_out。

使用seaborn中histplot直方图，对sms\_opp\_head\_unique\_count进行统计可视化，可以发现大部分短信的接收端的前缀码集中在1~5个号码之间。

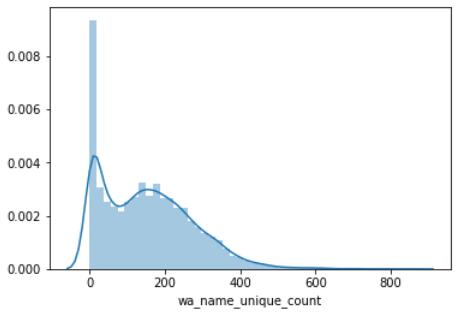
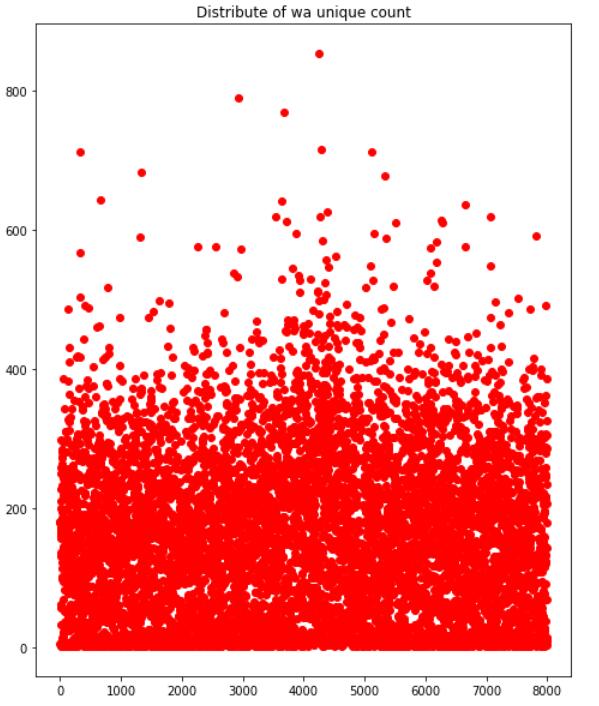


* wa:

关于wa模块，从中提取6个特征，分别为wa\_name\_unique\_count、wa\_name\_count、visit\_cnt、visit\_dura、up\_flow、down\_flow。

其中在visit\_cnt、visit\_dura、up\_flow、down\_flow几个特征集中在分别细化并计算他们的标准偏差、最小最大、中位数、平均、总数，这里也使用道聚合函数agg的功能，同时执行统计的操作。

关于每个uid访问不同网站总数的分布，可以观察到大部分都均匀分布在0~450之间的数量，图中有部分较突出的数据点可能为异常点或是风险用户。

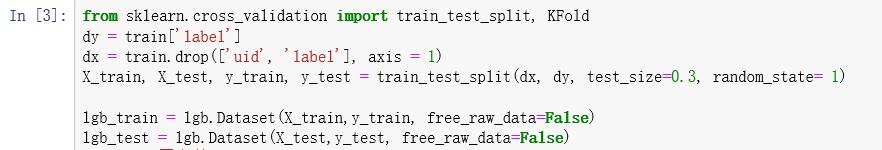


* LightGBM模型训练:

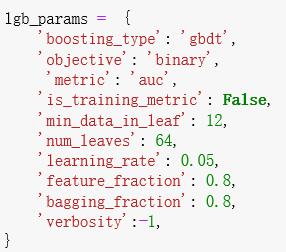
本次使用的分类是基于lightGBM构建的模型。LightGBM 是一个梯度 boosting 框架，使用基于学习算法的判定树。它可以说是分布式的，高效的。

* 调整参数:

为了得出较准确的模型参数，透过sklearn中的train\_test\_split将训练集中的数据再分成训练集跟测试集，再透过反复的交叉验证得出较好的结果并记录参数值。



首先，给定初始的参数值，具体在交叉验证后再做调整。



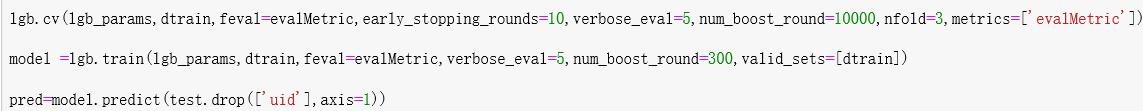
lightGBM使用的是leaf-wise的算法，因此在调节树的复杂程度时，使用的是num\_leaves而不是max\_depth。所以我透过loop循环语句，分别带入不同数目的num\_leaves以及max\_depth进行计算，将得出最好的结果的参数计入我们训练模型的参数中，其中，在模型的损失函数上metrics使用的是auc。代码如下:



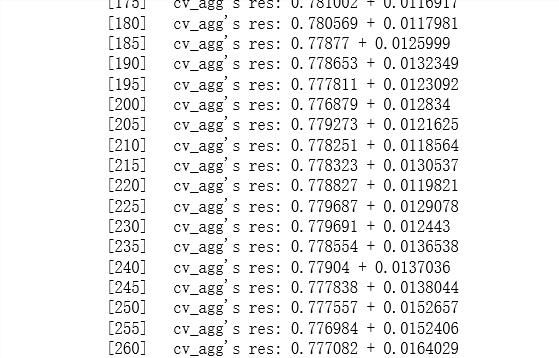
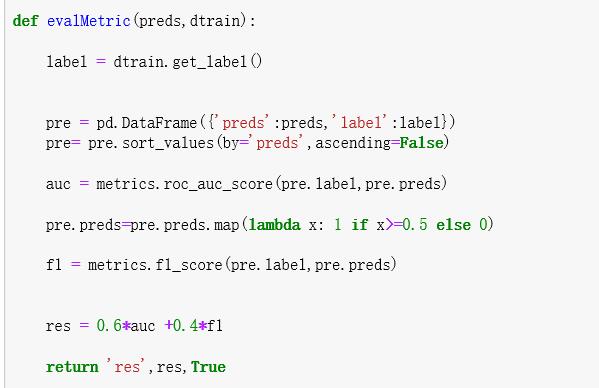
* 模型训练:

分别进行两次训练，第一次在给出初始化参数构件模型并预测；第二次在调整参数之后再进行训练查看是否有提高准确率。

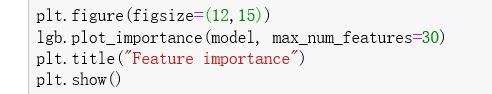
基本上lightGBM给出的训练语句为

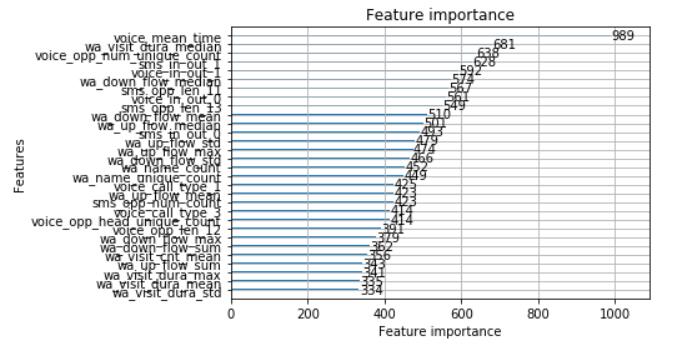


在feval参数中使用我们自己定义的函数evalMetric，在这个函数中计算本次项目的评定标准score=0.6×auc+0.4×F1，并用这个标准来训练我们的模型，并且下图也给出我们训练每轮迭代的结果。



可以根据以下代码来查看我们训练模型时特征的重要性的可视化，方便我们做特征的删减。





1. 项目排名、结果:



1. 项目心得与改进处:

本次项目是基于移动网络通讯行为的风险用户识别的实践，透过对实际案例的操作，帮助我更快掌握数据挖掘应用的方法以及增加本身对项目的分析和手作能力。

主要使用python中的机器学习模块sklearn、lightGBM等等，一些对数据操作的模块pandas、numpy，以及数据可视化操作matplotlib、seaborn。

项目主要分成两个部分完成，分别为数据整理、特征提取和分类模型的训练及预测。数据整理使用dataframe对象并在该对象上构建特征，这部有参考课件或相关案例等学习一些特征提取做法和想法，在构建训练模型时主要使用基于lightGBM的判定树算法，并在该模型上进行参数调整。

虽然最后顺利完成本次案例，但仍碰到些许困难并觉得有多处地方可以改进以及学习。首先，对lightGBM并未了解透彻，其中包括在构建该模型时所调用函数中并没有了解所有的参数以及作用，所以部分使用默认参数值，也参考多数现有案例的方案，导致降低了调参过程的效率以及准确性；第二，由于直接调用现有的函数训练，所以对算法实际的工作情况只有粗略的了解。这些部分会在课后自行学习和练习。