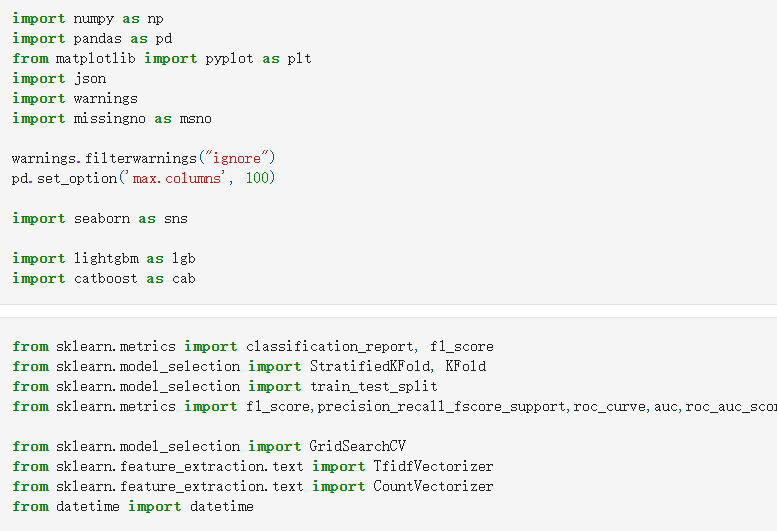


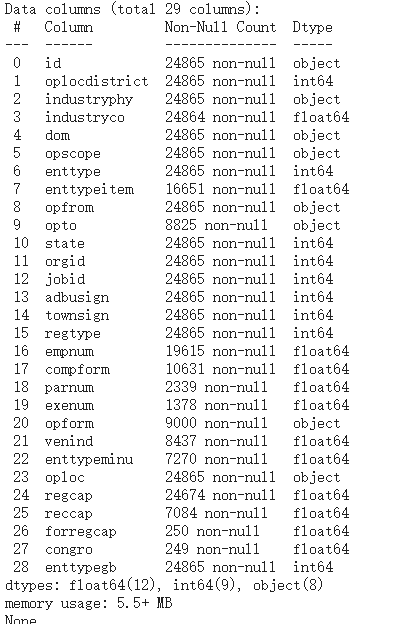
**企业非法集资风险预测**

|  |  |
| --- | --- |
| 姓名 | 班级 |
| 钱冠琦 | 机械91 |
| 严鑫茹 | 自动化93 |
| 刘勇 | 经济统计91 |
| 路逸诗 | 经济统计91 |

1. 实验目的
2. 掌握基本的Python数据探索步骤与技巧
3. 掌握并学会应用Python进行数据预处理的基本操作
4. 学会用Python实现分类模型实现训练与预测的功能
5. 提升自己。
6. 实验过程
7. 载入数据，并对数据进行初步了解分析，可以采用协方差分析的方法来选取最终要的提取数据。以下为本实验数据处理和训练与预测所需要载入的包，没有安装的包可以在命令行pip install进行安装，少部分包下载过慢可以使用清华大学tuna开源的镜像网站。



1. 数据处理。对不同的数据集可以采用不同的数据处理方式。
   1. 、对于base\_info，首先需要剔除“空列“，即没有数据或者数据太少的列，在这里，我们认为数据如果小于总id数目的5%就认为该列数据无效；然后需要对员工数量计数，处理时间获得企业经营时间，对经营范围进行字符串处理，对一些其他特征进行计数等，综上输出处理后的base\_info特征。



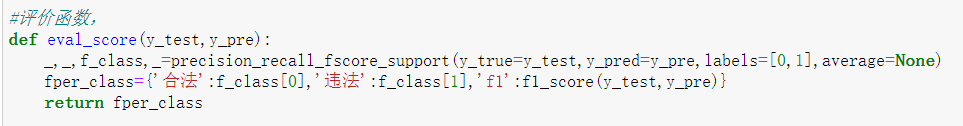
* 1. 、对于annual\_report\_info，同样需要剔除“空列“，按照id进行分组，对各组内分别分析特征Max,Sum等，并根据目标特征重新设置索引。
  2. 、对于news\_info，需要计算新闻距离现在过去了多少天，对新闻发布信息进行特征分析并reset\_index，之后重命名列columns，其中我们用数字123代表中立、积极与消极。
  3. 、tax\_info与annual\_report\_info大同小异，都需要按照所需要的特征groupby(‘id’)，并重新设置索引。
  4. 、change\_info需要把bgrq转换为可以使用的int类型，其他与以上步骤相同。
  5. 、other\_info需要进行计数

1. 数据拼接

通过pandas.merge拼接以上预处理得到的数据，并输出训练集和测试集，并输出训练集和测试集所有特征的信息。

1. 模型搭建

首先需要编写评价函数，通过参考本实验最后模型的评测依据，我们对各个评价指标进行了学习调研。这里可以采用库函数sklearn.metrics.precision.recall\_fscore\_support:计算每个类的精度、召回率、F 度量和支持度。详见官网document文档。



参数：

* 1. 、 precision:

精度是 tp / (tp + fp) 的比率，其中 tp 是真阳性的数量，fp 是假阳性的数量。

精度是分类器不将负样本标记为正样本的能力。

* 1. recall:

召回率是 tp / (tp + fn) 的比率，其中 tp 是真阳性的数量，fn 是假阴性的数量。

召回率直观地是分类器找到所有正样本的能力。

* 1. fscore:

F-beta 分数可以解释为精度和召回率的加权调和平均值，

其中 F-beta 分数在 1 时达到其最佳值，在 0 时达到最差分数。

F-beta 分数权重通过一个 beta 因子来决定召回率和精确度彼此间重要性的权值

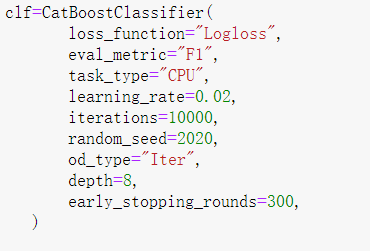
beta == 1.0 意味着召回率和准确率同等重要。

* 1. support支持度是每个类在 y\_true 中出现的次数。

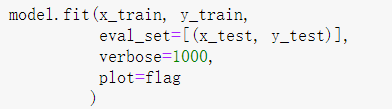
1. 模型训练

通过查阅资料和学习，我们选取了较为简单基础的两种训练方法，并调用了相关的包，我们采用了CatBoostClassifer训练以及LGBMClassifier训练，这部分由于能力和时间有限，我们无法对其内部原理进行比较全面的了解，只能采用“黑盒“思想暂时先完成任务。

以CatBoostClassifer为例。通过搭建模型调整参数可以建立我们的模型，模型中的参数调整是十分重要的，我们参照官网文档的叙述并通过调试采用了几个典型值。至于random\_seed可以随意取值，在学习到这一块时，我发现了一个很有趣的概念——伪随机数，也使得我对使用Python进行数据训练与分析背后的计算机运行原理有了初步的理解。



之后又进行了CatBoostClassifer.fit去训练我们搭建的模型，这一块函数的参数很难理解，而且还很多，最终只能选择了几个可以理解的参数进行调整，替他参数采用默认值。这一块的参数调整时根据最后的评价函数的fscore来调整的。



最终输出各个id评价结果，并保存到entprise\_submit.csv之中，并输出score mean

1. 反思

通过本次实验我们的收获很大，本次实验中我们夯实了基础，虽然学过python但从未进行过项目实践，这次的大作业就是一个良好的开端，我们掌握了基本的Python数据探索步骤与技巧，并学会应用Python进行数据预处理的基本操作。其次，我们见识到了python各种库函数和外部包功能的强大，基本可以说什么都可以做，果然全球开源工作者的力量是伟大的。我们学会用Python实现分类模型实现训练与预测的功能，并初步熟悉了模型参数调整的步骤，在写代码遇到bug的时候，可以即使找出问题所在并改正。

本次实验还有很多不尽如人意之处，在一开始的时候我尝试写代码，但是效率非常低下，最终只完成了数据提取和处理部分，而且大家的时间也有限，在大家的共同商量下，我们决定复现一篇去年的项目，参考一些已有的代码，并作出适当调整，它们已在文件夹中列出，希望老师谅解。

尽管我们最终得到了结果，但是对函数具体的内部过程知之甚少，从课堂上了解到的先进的算法和模型赶到要用代码实现的时候就变得无比艰难，经过东拼西凑才勉强能运行，要让自己写现阶段更是不可能，而且我们的模型显然是还有很大改进空间的，可以对参数进行优化，对看不懂在干什么的参数可以先进性数学等相关知识的学习，之后再引入模型之中。

出现这样问题的原因我觉得可能是大家都是新手，而且主修课业考试压力实在是太重了，对于以后的学习，可以先把一些基础的知识和近点的数目过一过，代码复现一下，而且英语的阅读理解能力也很重要，在查阅官网的时候阅读障碍很大，如果希望进一步学习data science的话需要提高自身的英语水平。之后再通过网络寻找相关项目的训练。

1. 分工

编写代码：钱冠琦

搜集资料：钱冠琦，刘勇，严鑫茹，路逸诗

撰写报告：钱冠琦

1. 鸣谢

感谢老师，您辛苦了，祝您工作顺利，阖家欢乐。