Project3-Adult Census Income

1. **Introduction:**

   本次project的工作是通过所提供的训练集拟合出合适的模型，并应用于所提供的测试集。具体问题为训练模型预测居民收入是否超过$50K/年。该数据集从美国1994年人口普查数据库抽取而来，其中类变量为年收入是否超过$50k，属性变量包含年龄，工种，学历，职业，人种等重要信息，值得一提的是，14个属性变量中有7个类别型变量。在本次project中，所训练的模型为两种，一种是线性逻辑回归模型，另一种是非线性梯度上升模型。

2. **Data Preprocessing:**

   首先检验所提供的traindata.csv中数据是否有缺失，结果为没有数据缺失。然后我们将14个属性分为两类，一类是其字段值连续的，另一类其字段值是离散的。对其中数据为离散的，我们采用独热码进行编码。而为什么不考虑运用labelencoder对其进行编码，是考虑到labelencoder对离散数据进行编码后不能保证数据间距离的相同的，这与所探究问题需要的模型不契合。而对其中数据为连续的，我们采用标准正态的方式进行归一化，目的是为了消除数据中的奇异值，是模型拟合更为契合，同时也加快了训练的收敛性。之后我们利用交叉分层验证的方式对训练集进行了划分。最后调用SGDClassifier和GradientBoostingClassifier进行拟合，并将训练好的两个模型通过joblib模块提取至save文件夹下，分别命名为lr.pkl和gbc.pkl。这使得对于任意后续提供的测试集，我们都可以通过只运行一个test.py的测试文件来得到预测结果，这也对训练好的模型进行了一定的保护。

3. **Model Classifiers:**

   SGDClassifier: 模型每次使用一个样本来估计损失函数梯度。模型的学习速率会随着迭代地进行而减小

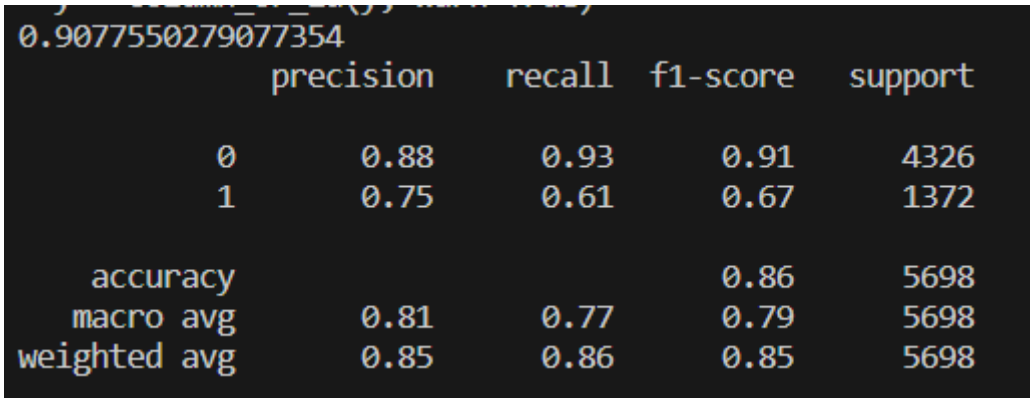
   优点： 效率高。 易于实施（有大量调整代码的余地）。 适用于大规模和稀疏问题上， **缺点**： 有许多超参数。 对数据的范围很敏感，需要对数据进行缩放/标准化

   GradientBoostingClassifier:根据当前模型损失函数的负梯度信息来训练新加入的弱分类器，然后将训练好的弱分类器以累加的形式结合到现有模型中。

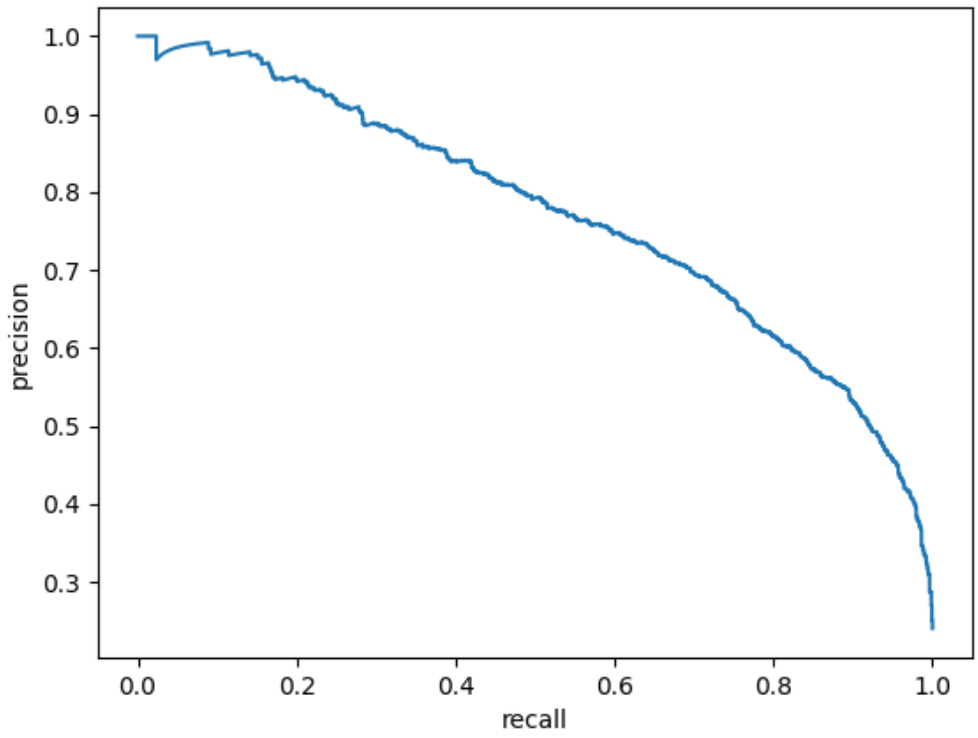
   优点：准确率高，**缺点**：需要仔细调参，通常不适用于高维稀疏数据

4. **Model Evaluation:**

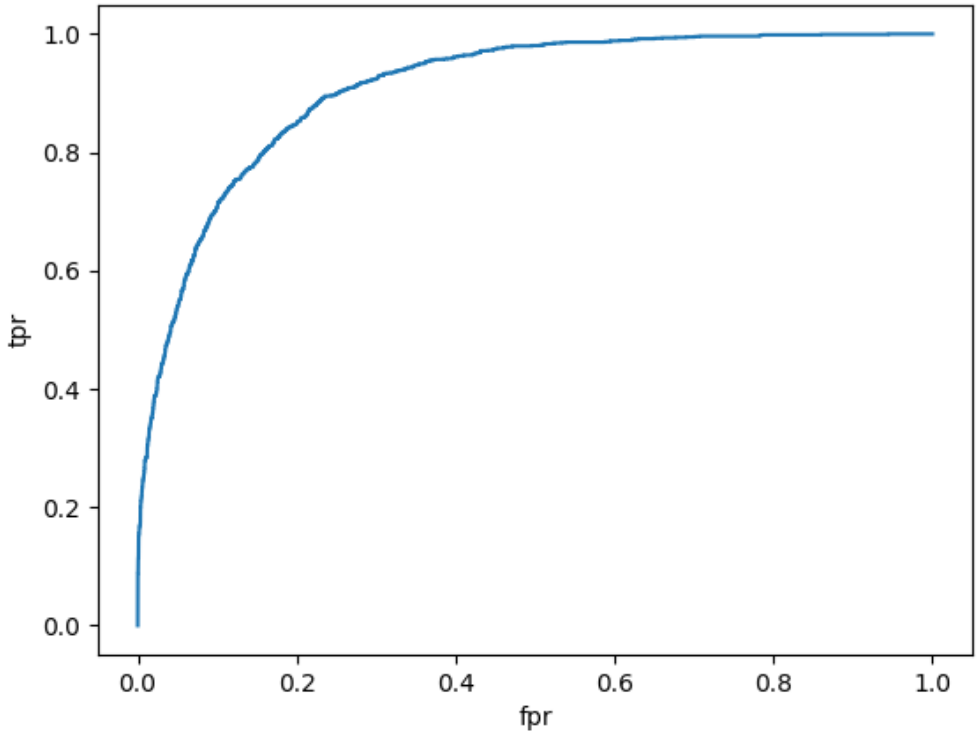
   SGDClassifier:



   PR曲线:



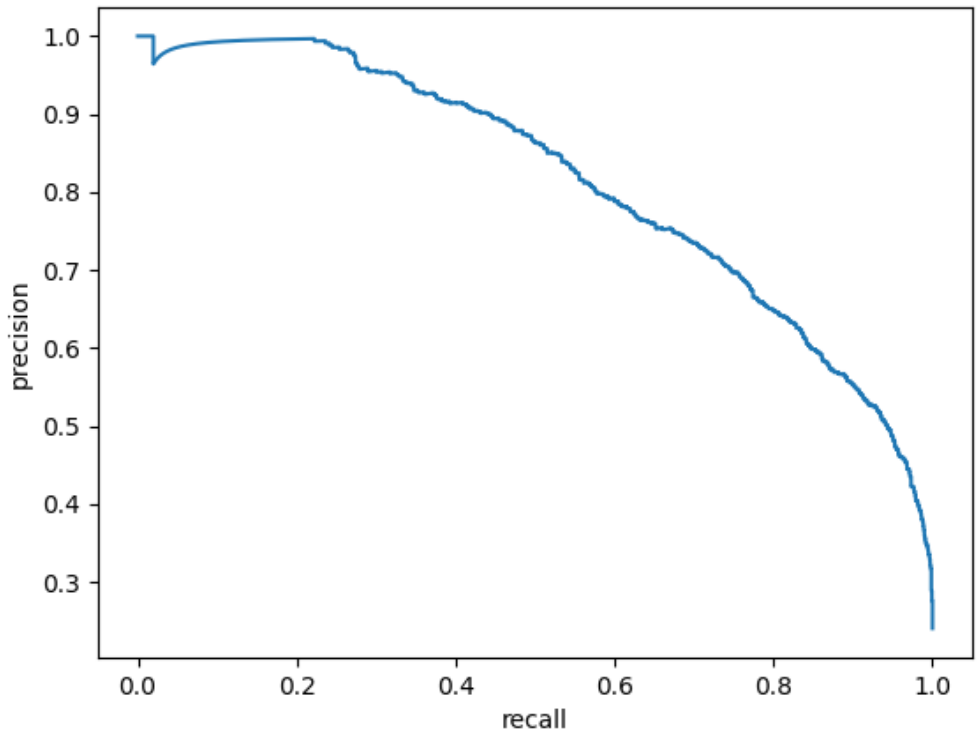
ROC曲线:



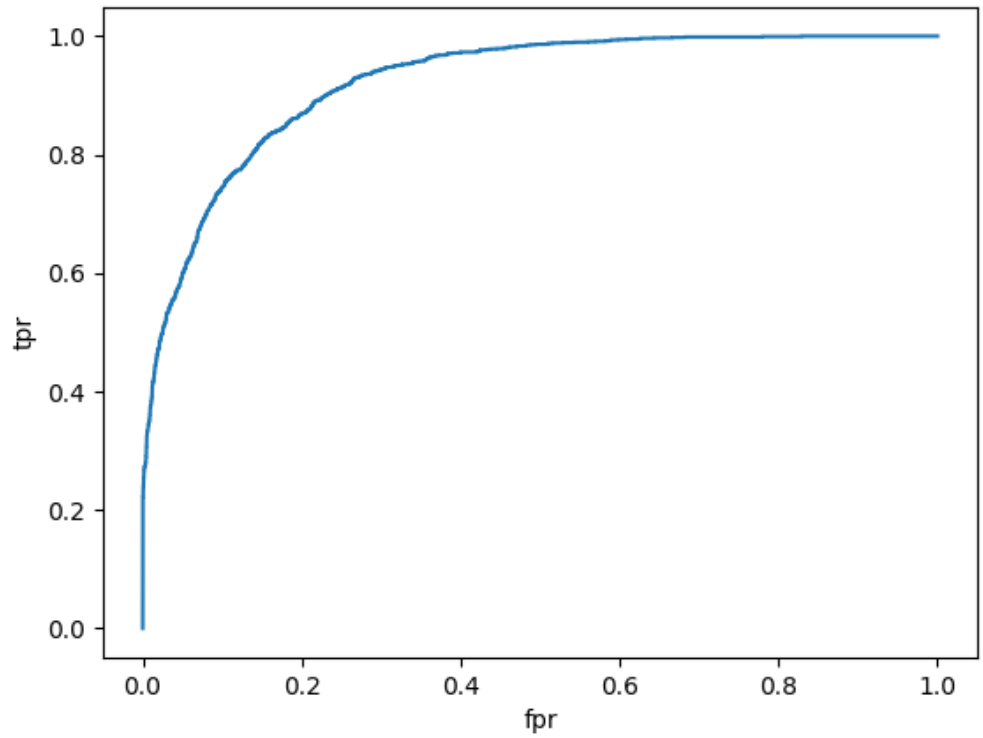
GradientBoostingClassifier:

```
0.9277125799794853
              precision    recall  f1-score   support

           0       0.90      0.94      0.92      4326
           1       0.79      0.66      0.72      1372

    accuracy                           0.87      5698
   macro avg       0.84      0.80      0.82      5698
weighted avg       0.87      0.87      0.87      5698
```

PR曲线:



ROC曲线:



分析：两个模型都起到了不错的拟合效果，并且第二个非线性模型在各方面优于第一个线性模型。第二个模型的auc值比第一个模型的更大，准确率更高，但差别并没有非常明显。两个模型的准确率都维持在85%-90%，仍有空间提升。

5. **Results and Discussion:**

   对于测试集的模型预测结果存储在testlabel.txt中，在test.py文件中我们可以设置输出的预测结果为第一类模型还是第二类模型。值得一提的是，我们需要对测试集数据进行独热码编码的时候与训练集的维数对齐，而本project中我们将多于的维数全部舍弃即设置字段值为0。

6. **Limitations and Future Work:**

   由于时间关系，事实上本次project中并没有涉及所有的模型，但在查阅资料的过程中，我们所采用的第二种集成学习的分类器的表现实际上是数一数二的，而采用逻辑回归线性模型是因为它是最基本的一个模型。在后续的完善工作中，我们会对其他模型也进行相应的探究。

7. **Conclusion:**

   在这次project中，探究了所学机器学习内容在实际生活中的应用，这使得我对所学内容有了更深层次的理解和领悟，同时在过程中对sklearn库也有了更多的了解，学会了许多常用且实用的函数，也学会了数据处理分析方面的知识。这也让我对数据处理分析，机器学习方面知识产生了想要探索更多的兴趣。最后，在本次project中没有尝试的模型，后续都会进行测试，通过本次project对这些所学常用的模型有了更深的体会。

8. **References:**

   Python数据分析—成年人收入水平预测