Requires Changes

**还需满足 3 个要求 变化**

同学你好，第一次的项目提交要注意的细节比较多，认真审阅此次review修改一下模块:

* 算法流程的修改
* sample\_size的代码修改
* 模型通俗叙述的修改

**探索数据**

**学生正确地计算了下列数值:**

* **记录的数目**
* **收入大于50000美金的人数**
* **收入小于等于50000美金的人数**
* **收入大于50000美金的人数所占百分比**

非常好，你的计算完全正确。

**准备数据**

**学生正确地对特征和目标实现了独热编码。**

很不错的独热编码，replace函数的用法非常灵活，你可以这样写来简化你的代码:

income = income\_raw.replace(['>50K', '<=50K'], [1, 0])

**评估模型表现**

**学生正确的计算了简单预测的准确率和F1分数。**

**学生解释了选择这几个模型的原因，并说明了每一个模型的优缺点。**

Logistic Regression逻辑回归是可以处理非线性问题的, 但是需要先对特征作非线性变换,使用kernal trick 看看这里:<https://www.zhihu.com/question/29385169>

这个[知乎回答](https://www.zhihu.com/question/21094489" \t "_blank)对svm做了通俗的讲解，对形成基本的知识体系有很好的帮助。

如果想深入支持向量机可以学习Free Mind系列的博客:<http://blog.pluskid.org/?page_id=683>

**学生成功的实现了一个监督学习算法的流程。**

不错的算法流程编写，但是有几个细节仍需注意。

**warning1**

注意注释的要求然后得到对前300个训练数据的预测结果，而不是对sample\_size大小的数据预测结果。所以你需要对这一行代码predictions\_train = learner.predict(X\_split)以及相关联的代码进行相应的修改。

**warning2**

在fbeta\_score中你设置了average参数

fbeta\_score(y\_train, predictions\_train,average='macro',beta=0.5)

通过查阅文档对average参数的说明:  
**This parameter is required for multiclass/multilabel targets. If None, the scores for each class are returned. Otherwise, this determines the type of averaging performed on the data**  
这个参数经常用于用于多类别分类或者多标签分类问题上，对于二分类问题，最好是设置average = None。虽说在这个问题上，average = 'binary'跟前者效果是一样的，但是设置average = None能够自动处理one-hot后的label，而且避免后面使用fbeta\_score时，忘记设置average = binary这种情况，使得结果不一致的情况。

**学生正确的实现了三个监督学习模型，得出了模型表现可视化的图表。**

注意samples\_1，samples\_10以及samples\_100必须为**相对应数据的大小**，而**不是**存储一个DataFrame类型的数据(这是造成在打印的时候出现大量输出的主要原因)。在对sample\_size做相应的修改后，别忘了在上一问的train\_predict函数中修改对应的代码。

**优化结果**

**在考虑了计算成本、模型表现和数据特点之后，学生选出了最好的模型并给出了充足的理由。**

**学生能够用清晰简洁的话来向一个没有机器学习或任何其他技术背景的人来解释最优模型的工作原理。**

明确这里的目的:  
机器学习模型一般都有这三个要点需要解释:

* 模型的基本流程．(你已经提到逻辑回归需要得到一个概率模型，但是你还应该提到逻辑回归中特征的权重这个概念)
* 优化目标．(你应该提及损失函数)
* 模型的训练是如何进行的 (可以提到数据的作用，机器学习一定要提到数据，最好能解释一个训练方法，如梯度下降)．  
  有关logistic regression更加具体的内容可参考[这里](http://www.cnblogs.com/sparkwen/p/3441197.html" \t "_blank)

**最终模型利用了网格搜索进行参数调优，至少挑战了一个参数，并且至少有三个可选值。如果模型参数不需要任何调整，学生需要给出明确的理由。**

很不错的调参代码编写。或许你需要调整logistic regression里面的C参数，最后结果的准确率或许有不错的提高。

**学生在表格中正确汇报了调优过后、调优之前以及基准模型的准确率和 F1 分数。学生把最终模型的结果与之前得到的结果进行了对比。**

**特征重要性**

**学生列出了他们认为对预测个人收入最重要的5个特征，同时给出了选择这些特征的理由。**

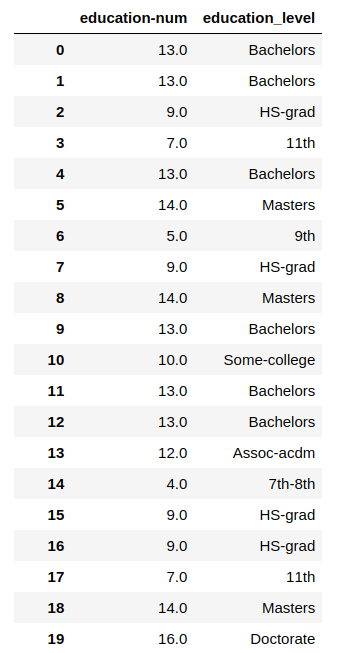
**学生调用了一个监督学习模型的 feature\_importances\_ 属性。此外，学生列出了这些重要的特征并讨论了这些特征的相同点和不同点。**

education-num不仅代表教育时长，而且它是education\_level的labelEncoding结果，某种程度来说也代表学习水平。

以下代码可以清楚解释这个原因，注意观察以下代码的输出值：

zip(list(data.education\_level.values), list(data['education-num'].values))

这里显示前20个数据：

[](http://upload-images.jianshu.io/upload_images/1132123-f122e7af8cd97f75.png?imageMogr2/auto-orient/strip%7CimageView2/2/w/1240)

**学生用最重要的5个特征建模并分析了和对比了改模型与问题五中的最优模型的表现。**

**对于问题八**  
你还可以从这几个方面入手来思考这个问题：

* 这个模型是不是不可以做[online training](https://en.wikipedia.org/wiki/Online_machine_learning)。如果不能，在实际生产中，训练结束之后把模型保存下来，就可以不停地使用，训练时间就不是考虑的范畴。
* 我从特征选择和特征降维的角度入手，把要丢弃的特征利用起来，在保证训练和测试时间(视降维算法而言)的同时在一定程度上也能有较高的精度。
  + 关于特征选择通俗易懂的[说明](http://machinelearningmastery.com/an-introduction-to-feature-selection/)，介绍了什么是特征选择，特征  
    选择可以解决什么问题，以及常用的特征选择的方法。
  + 接下来要介绍特征选择的[进阶版](http://machinelearningmastery.com/feature-selection-machine-learning-python/)，配合特征降维来使用，大大提高模型的性能
  + 配合sklearn来做[特征选择](http://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/)，简单明了
* 并且随着硬件以及算法的发展，模型的训练速度会不断提升，准确率才是我们评价一个模型好坏的最终标准。











