### 风控分析报告

### 建模思路

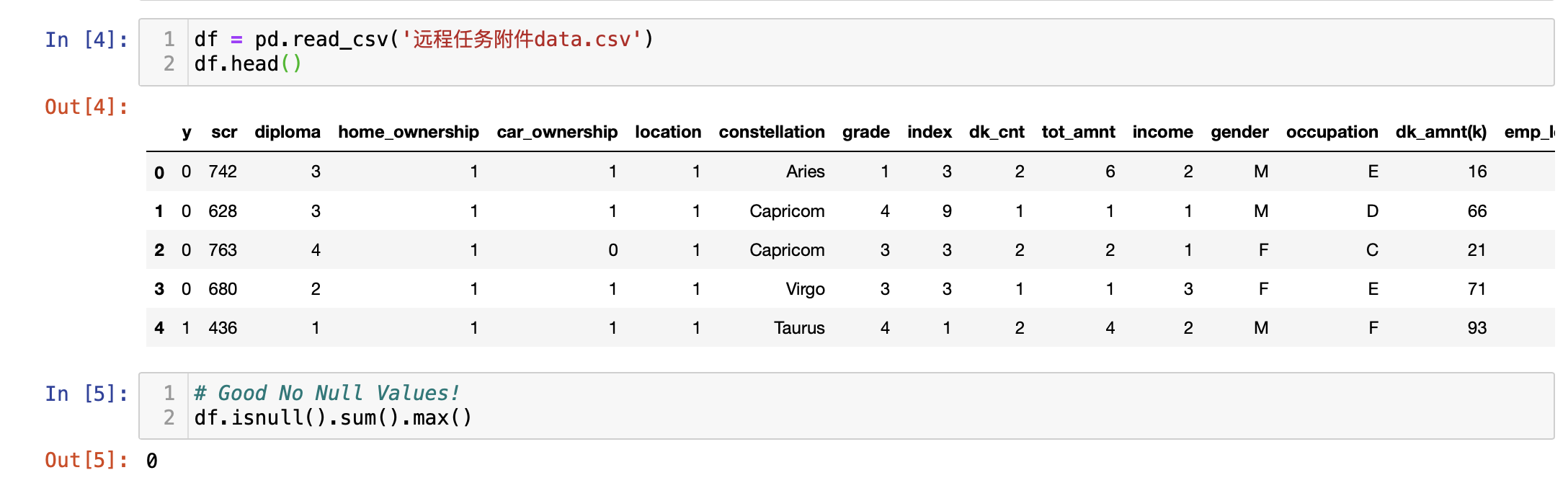
### 

1. **场景解析**

贷款申请人向平台申请贷款时，平台通过线上或线下让客户填写贷款申请表，收集客户的基本信息，这里包括申请人的年龄、性别、学历、贷款金额、申请人财产情况等信息，通常来说还会借助第三方平台如征信机构的信息。通过这些信息属性来做回归 ，生成预测模型，平台可以通过预测判断贷款申请是否会违约，从而决定是否向申请人发放贷款。

首先，我们的场景是通过用户的历史行为（如历史数据的多维特征和贷款状态是否违约）来训练模型，通过这个模型对新增的贷款人“是否具有偿还能力，是否具有偿债意愿”进行分析，预测贷款申请人是否会发生违约贷款。这是一个监督学习(supervised learning)的场景，因为已知了特征以及贷款状态（Y字段）是否违约（目标列），我们判定贷款申请人是否违约是一个二元分类问题，可以通过一个分类算法来处理，这里我们将用三种分类算法进行建模并选取最优的模型。

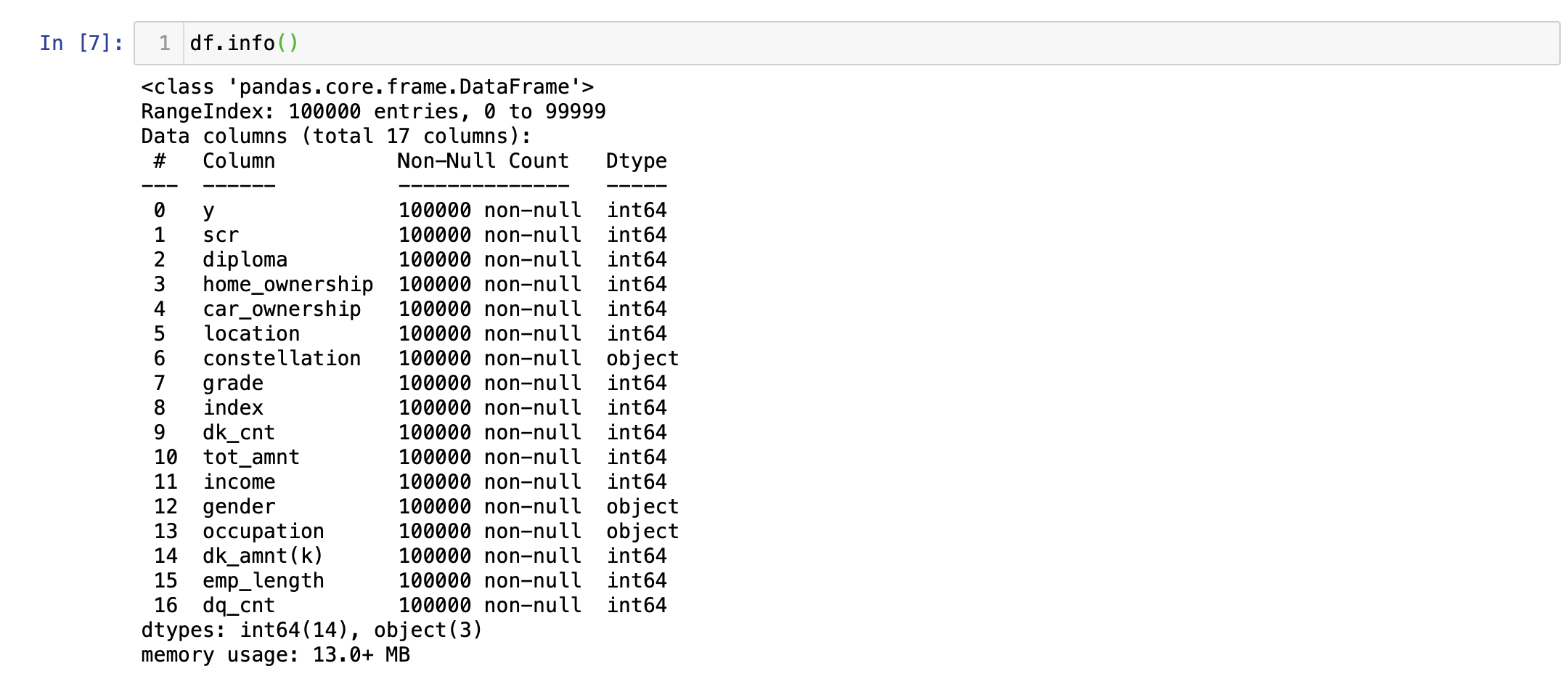
1. **数据预处理**
2. **缺失值处理**



我们首先通过pandas提供的read\_csv途径读取样本数据，进行简单的分析，发现本样本中并没有存在缺失值。

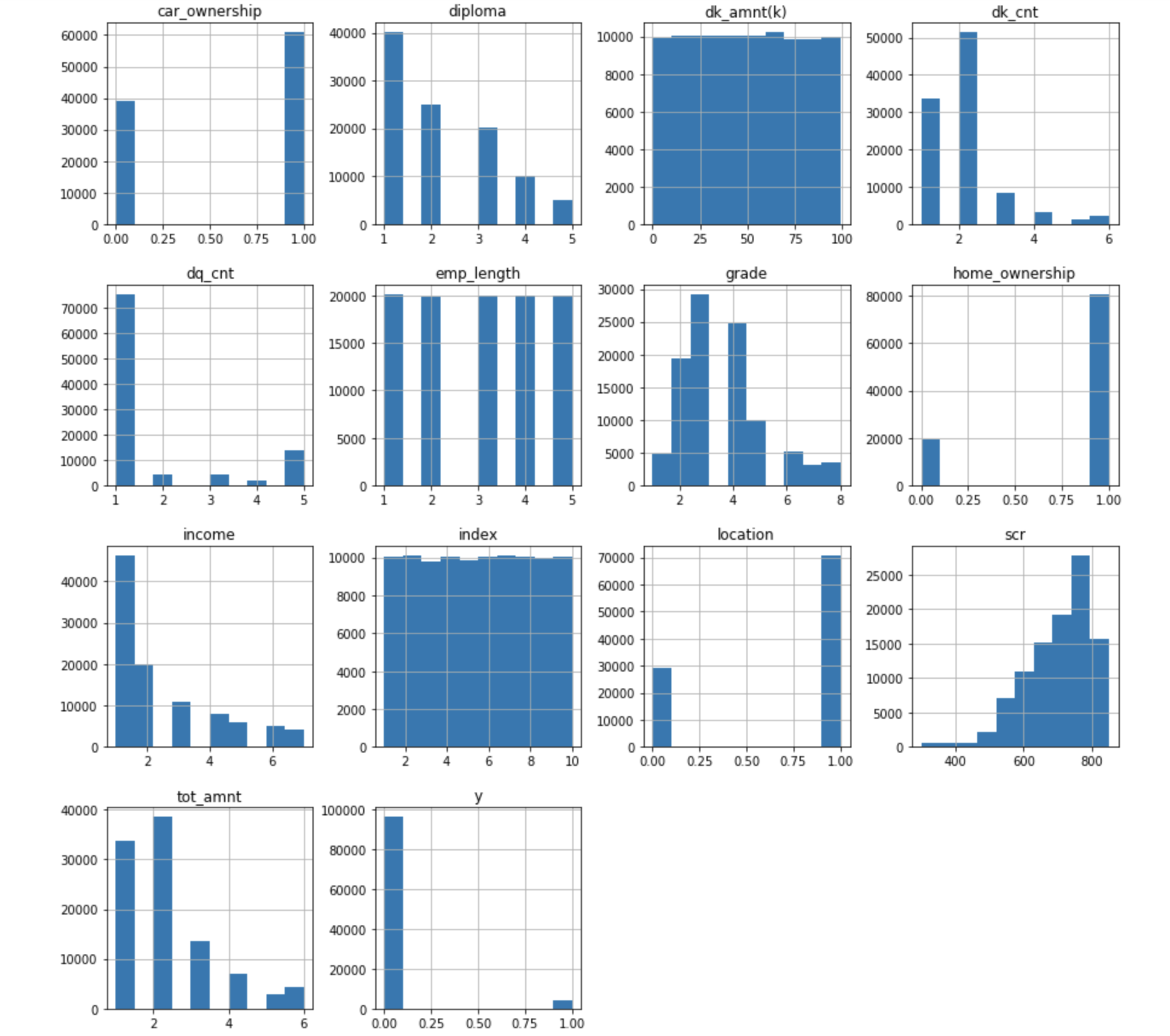
1. **变量的分布**

特征值的类型



我们可以看到样本中，int64类型的变量包括：y，scr，diploma，home\_ownership，car\_ownership等，object类型的数据有constellation，gender和occupation；

对于object的特征我们会在下一步进行one\_hot的编码。在这一步我们会观察数值型的变量的分布情况。



我们能看到在这些分布中，并不存在特别明显的异常值，所以我们保留原样本数据作为我们的分析数据

1. **特征工程**

在特征过程中，我们需要对特征值做一些相应的处理来达到以下的目的：

* 处理object变量，令所有的变量类型为数值型，从而复合训练模型的要求；
* 标准化特征值
* 特征选择：选择和目标相关性较大的变量；去除冗余特征；

1. **One\_hot 编码**

对object变量进行one\_hot 编码，获得全数值型特征值。



1. **标准化特征值**

调用scikit-learn模块preprocessing的子模块StandardScaler进行数据标准化处理



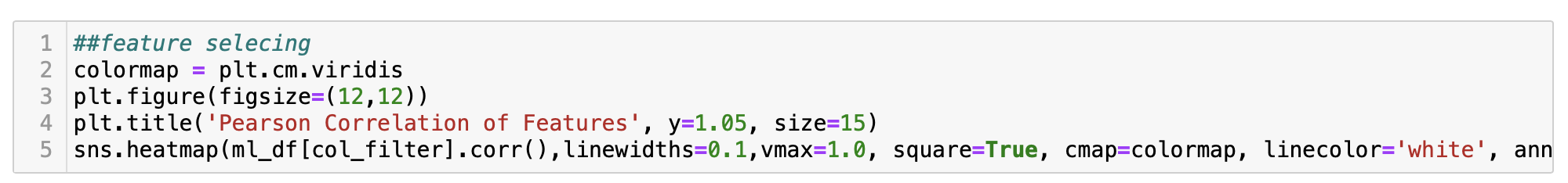
1. **特征选择**

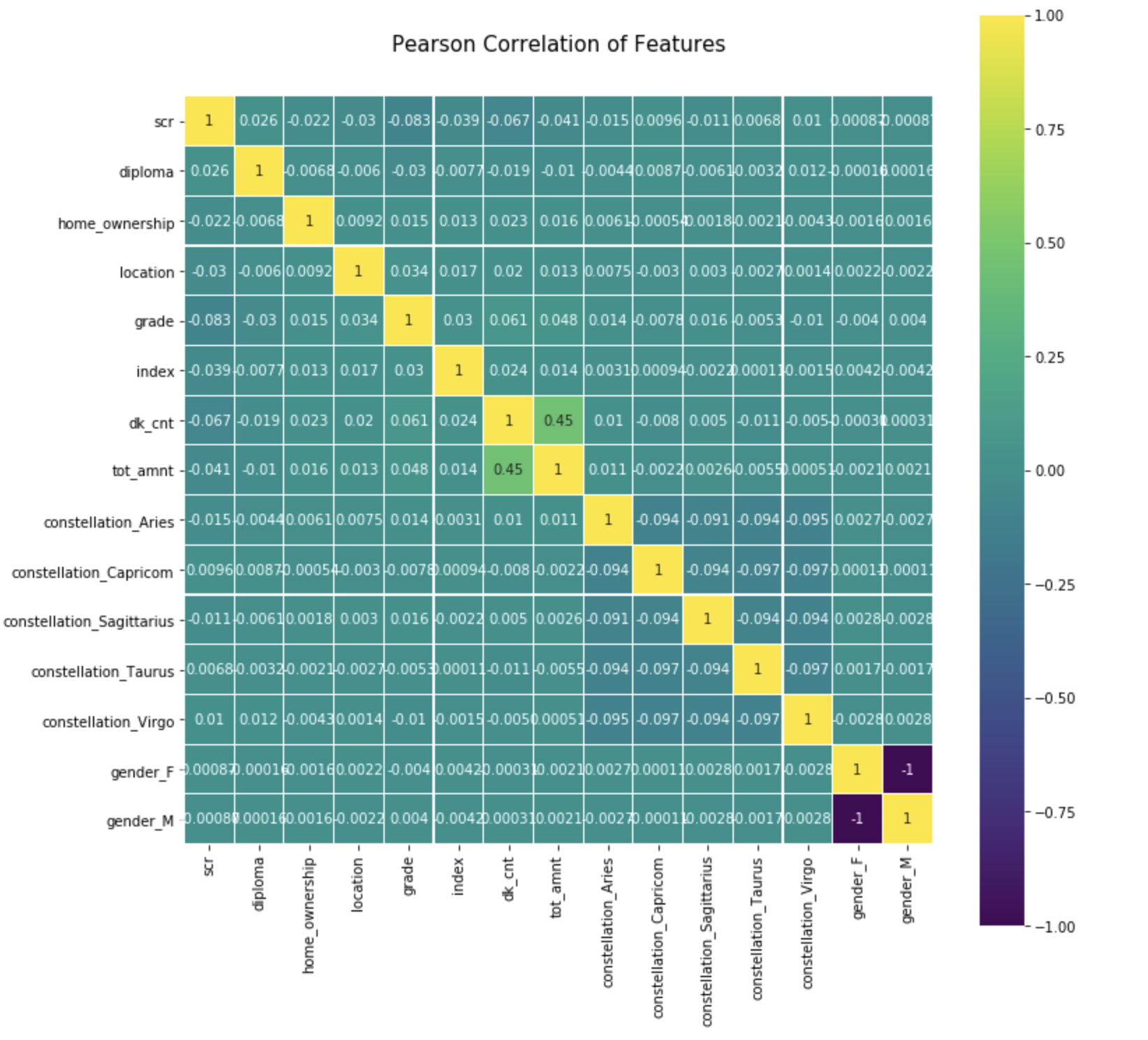
选择和目标相关性较大的变量，我们能看到原模型有34个特征值；



我们将采用的Wrapper方法，通过暴力的递归特征消除 (Recursive Feature Elimination)方法筛选15个与目标变量相关性最强的特征，逐步剔除特征从而达到 首次降维，自变量从34个降到15个。

接下来，我们观察一下15个特征项之间的相关系数，发现存在冗余项（高度相关），gender\_M 和 gender\_F存在高度相关，为冗余项，我们保留一项。







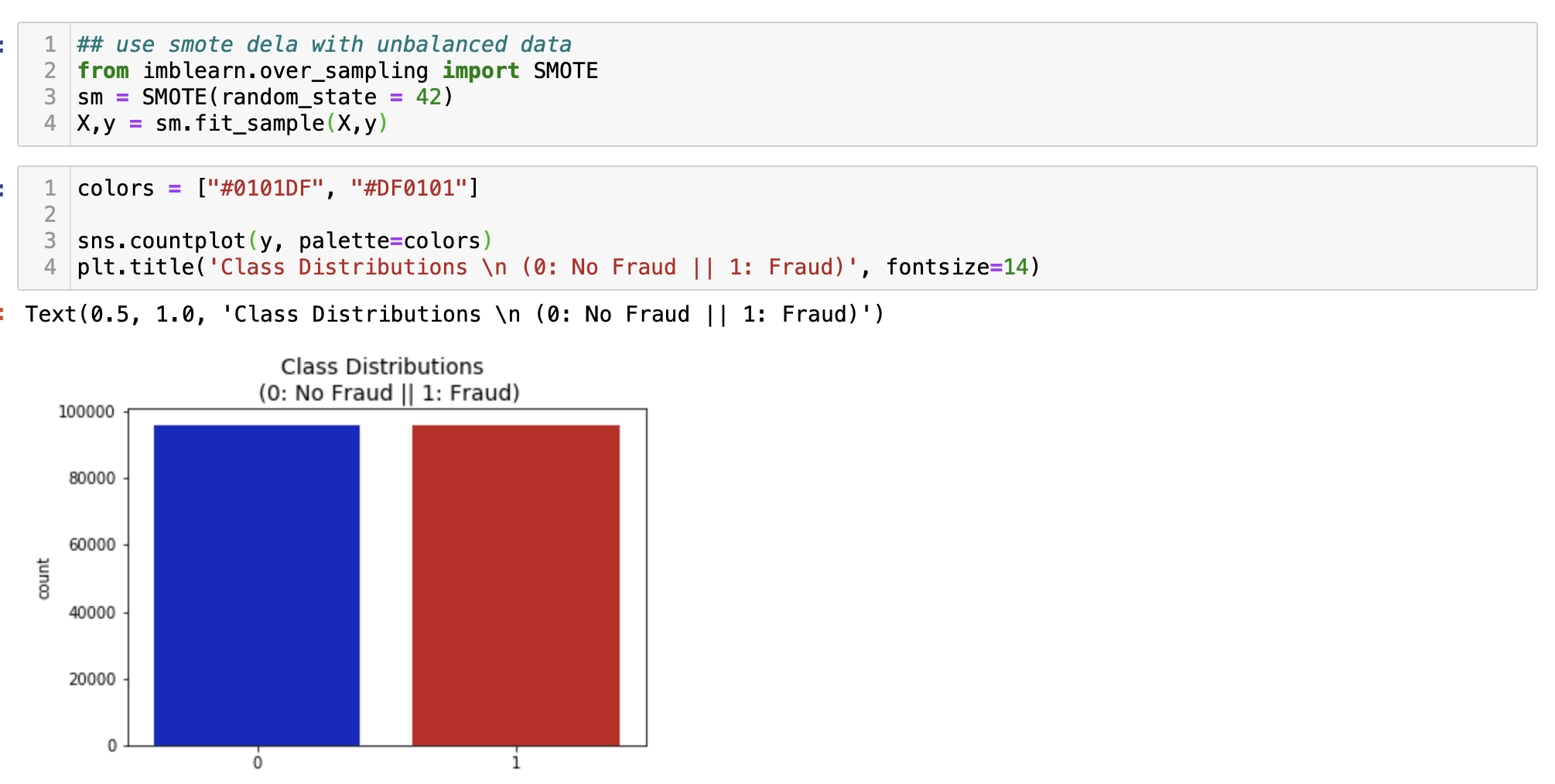
1. **模型的训练**
2. **处理样本不均衡（\*）**

由于项目要求需要运用全样本数据，所以之后的模型还是建立在全样本数据上，我会将采用更新的数据集来训练的模型放置在最后的总结与思考中。

在本次项目中，我们能发现样本数据集是不平衡的，这会影响到我们训练的模型。

我们可以看到非违约比例达到了96%，而违约比例只有4%。这会导致我们模型的预测正确率的有效性失效，模型只需要能够给出大概率的非违约目标就可以实现较高准确的预测能力。所以我会在项目中的总结部分，用平衡的数据集代替原样本数据。

我们将采用SMOTE方法来处理我们的非平衡的数据集。

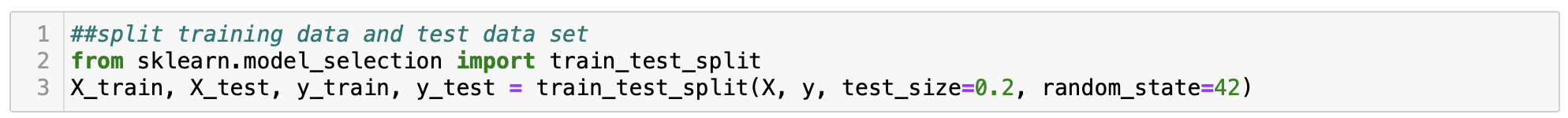


1. **构建分类器训练**

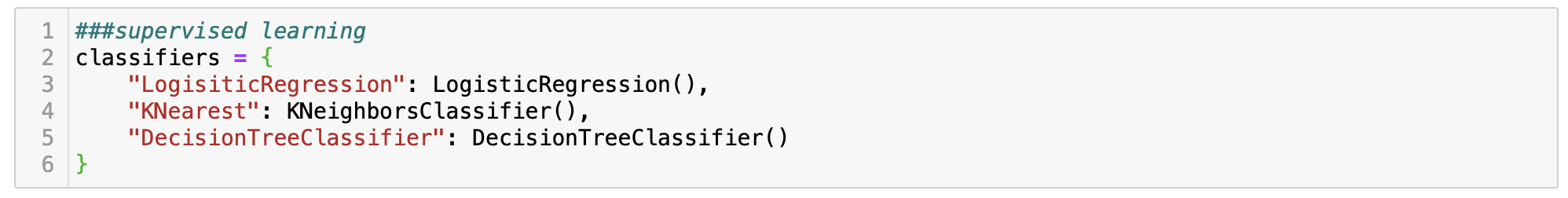
本次项目我们采用交叉验证法划分数据集，将数据划分为3部分：训练集（training set）、验证集（validation set）和测试集（test set）。让模型在训练集进行学习，在验证集上进行参数调优，最后使用测试集数据评估模型的性能。模型调优我们采用网格搜索调优参数（grid search），通过构建参数候选集合，然后网格搜索会穷举各种参数组合，根据设定评定的评分机制找到最好的那一组设置。

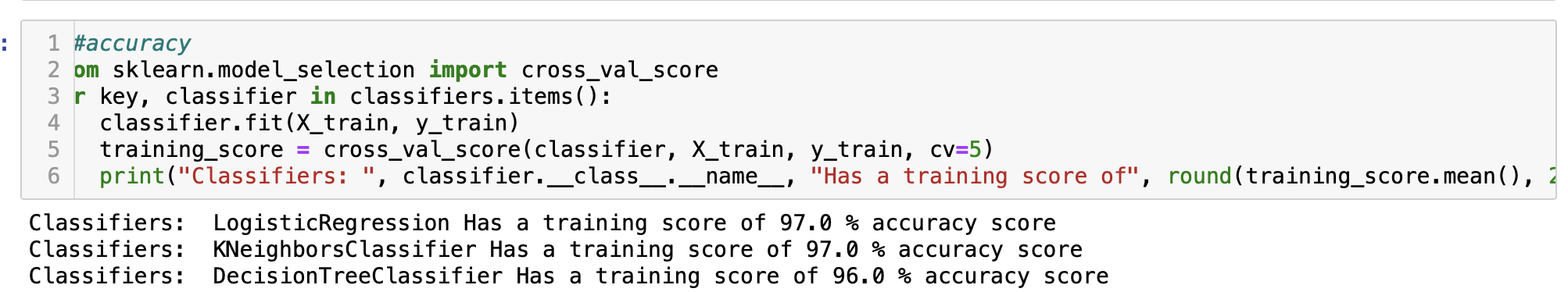
模型我们分别选择logistic、KNearNeighbors和decision tree来构建。

划分数据集：



训练三种模型并得到各自的训练准确度：

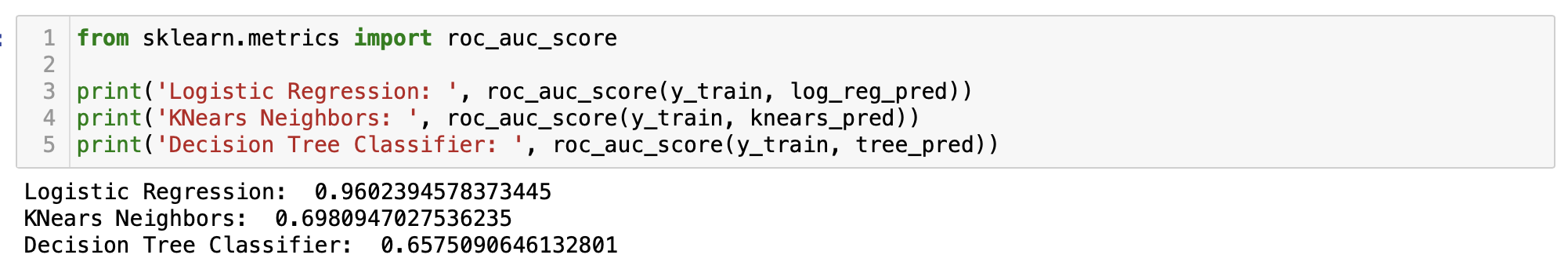


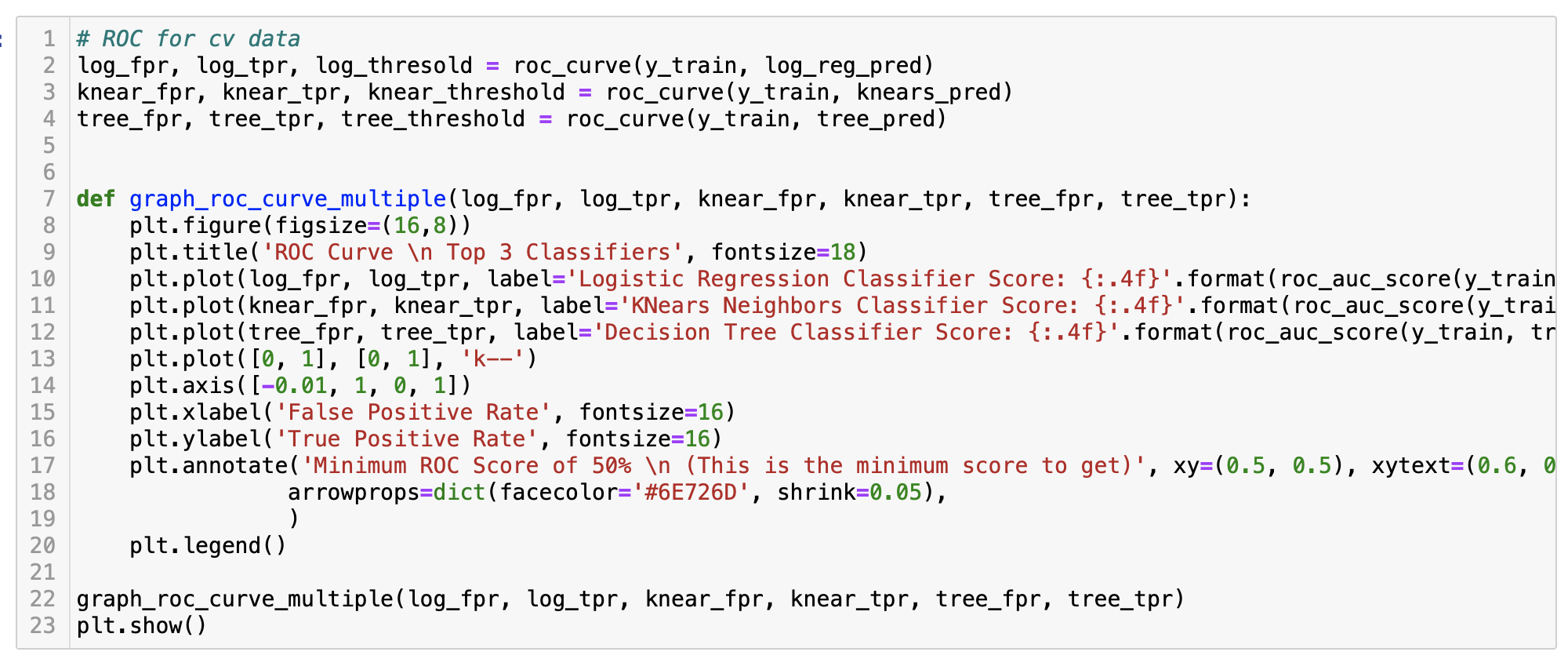


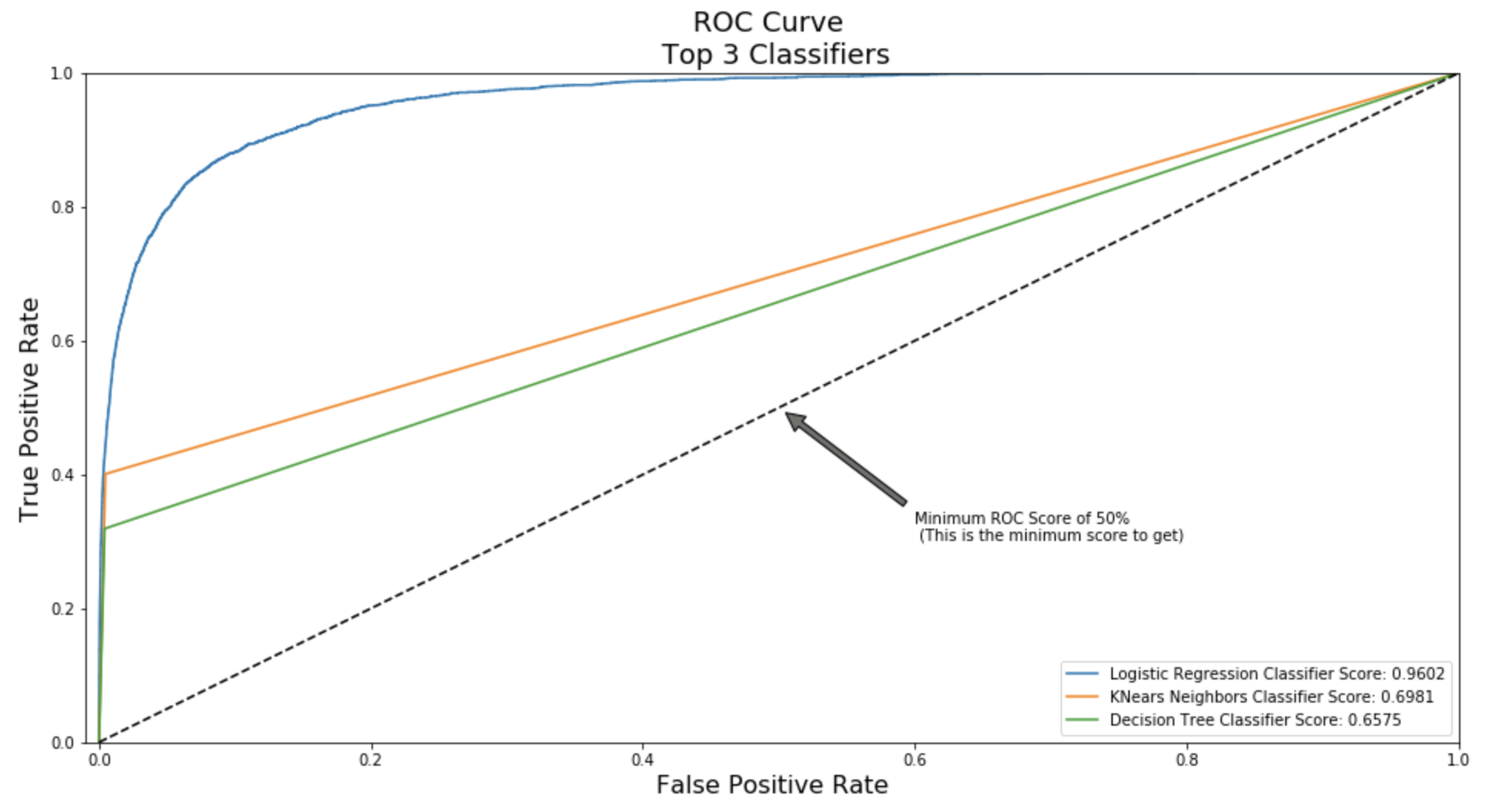
利用网格搜索调优参数来调整模型，选取最好的参数



利用交叉验证来测试三个模型ROC





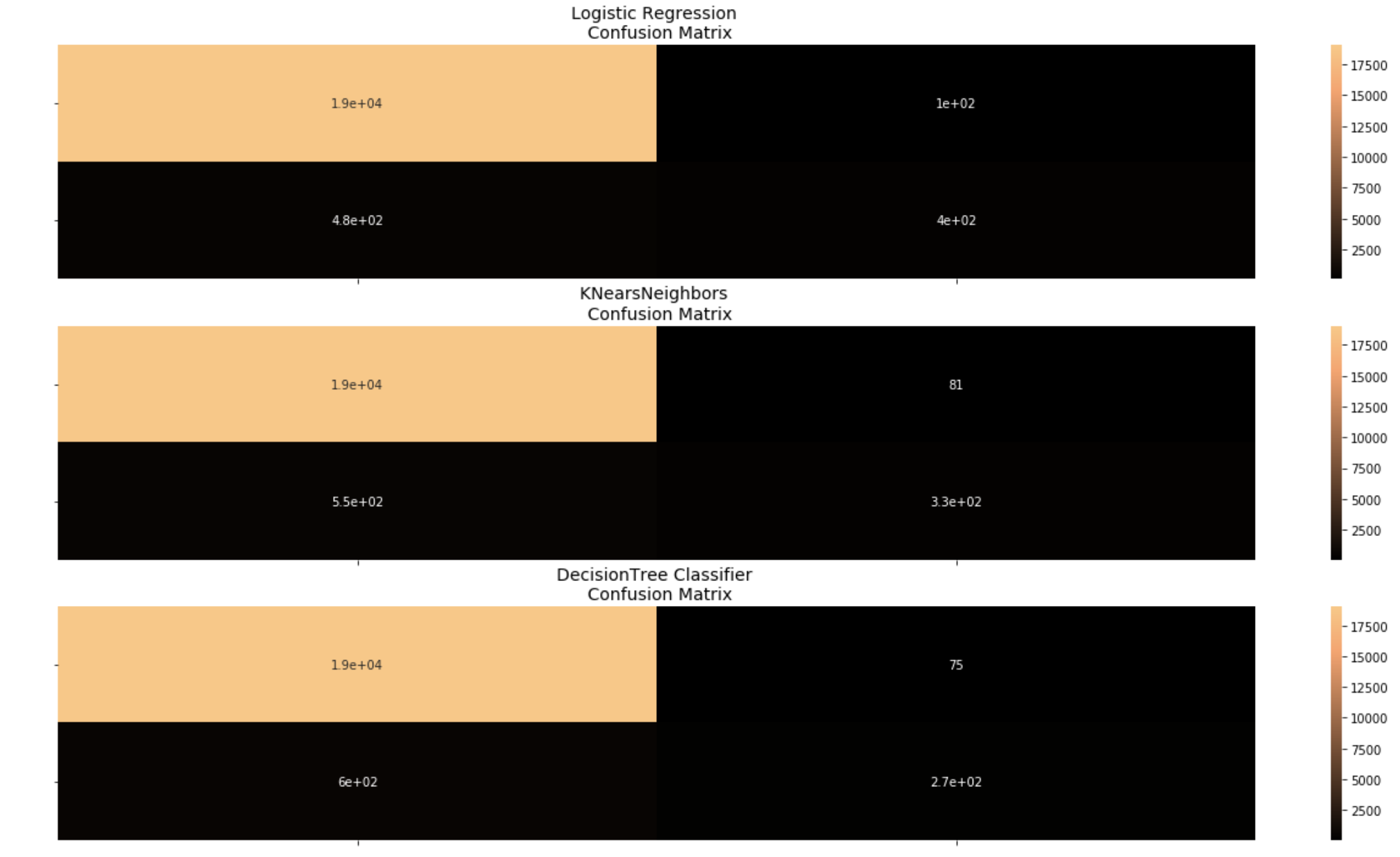


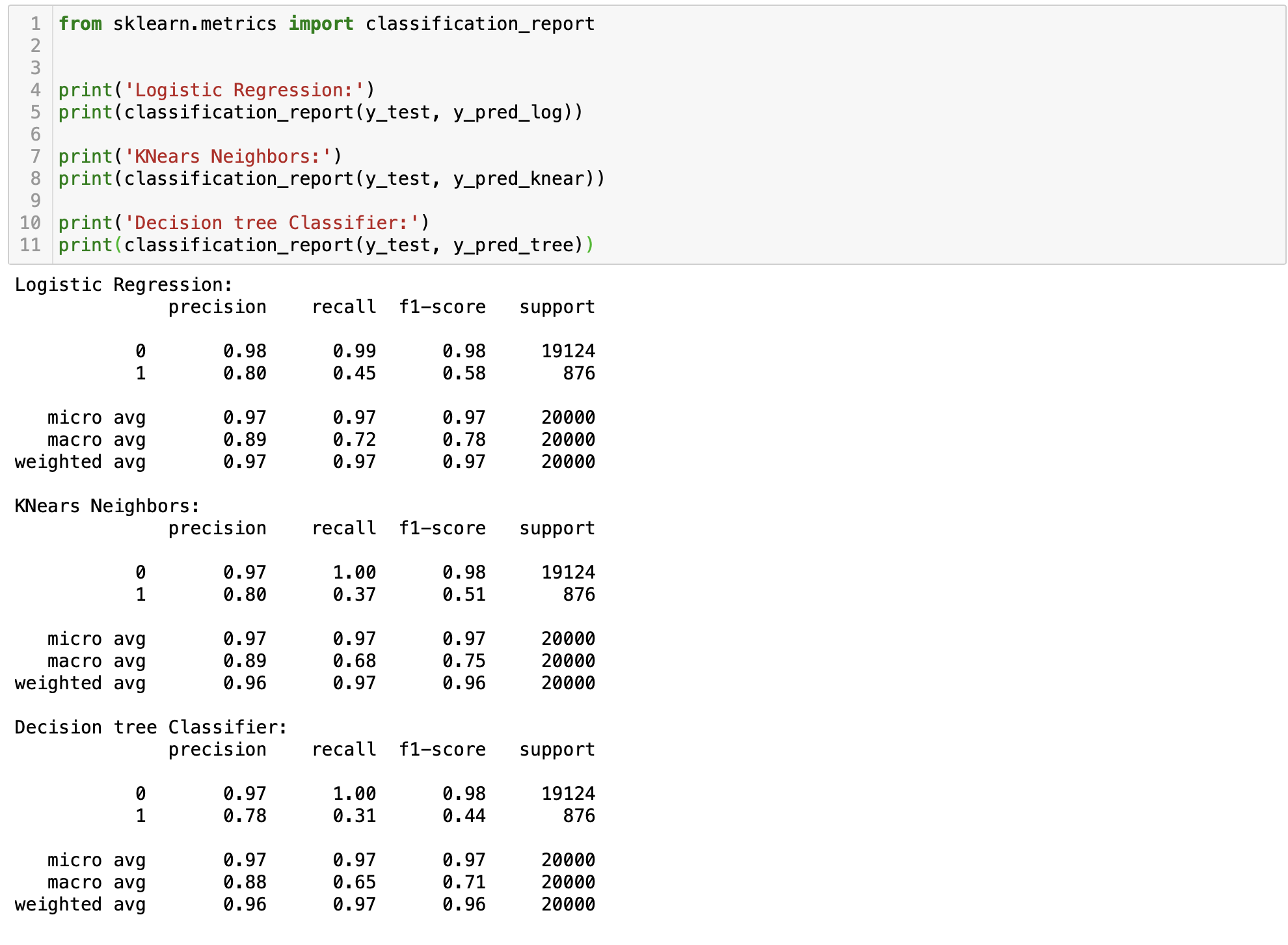
我们能看到在训练过程中，三种模型中logistical model在交叉严重中表现的最好。

1. **模型的测试**
2. **混淆矩阵**

将调优的模型在测试集上测试，得到对应的混淆矩阵

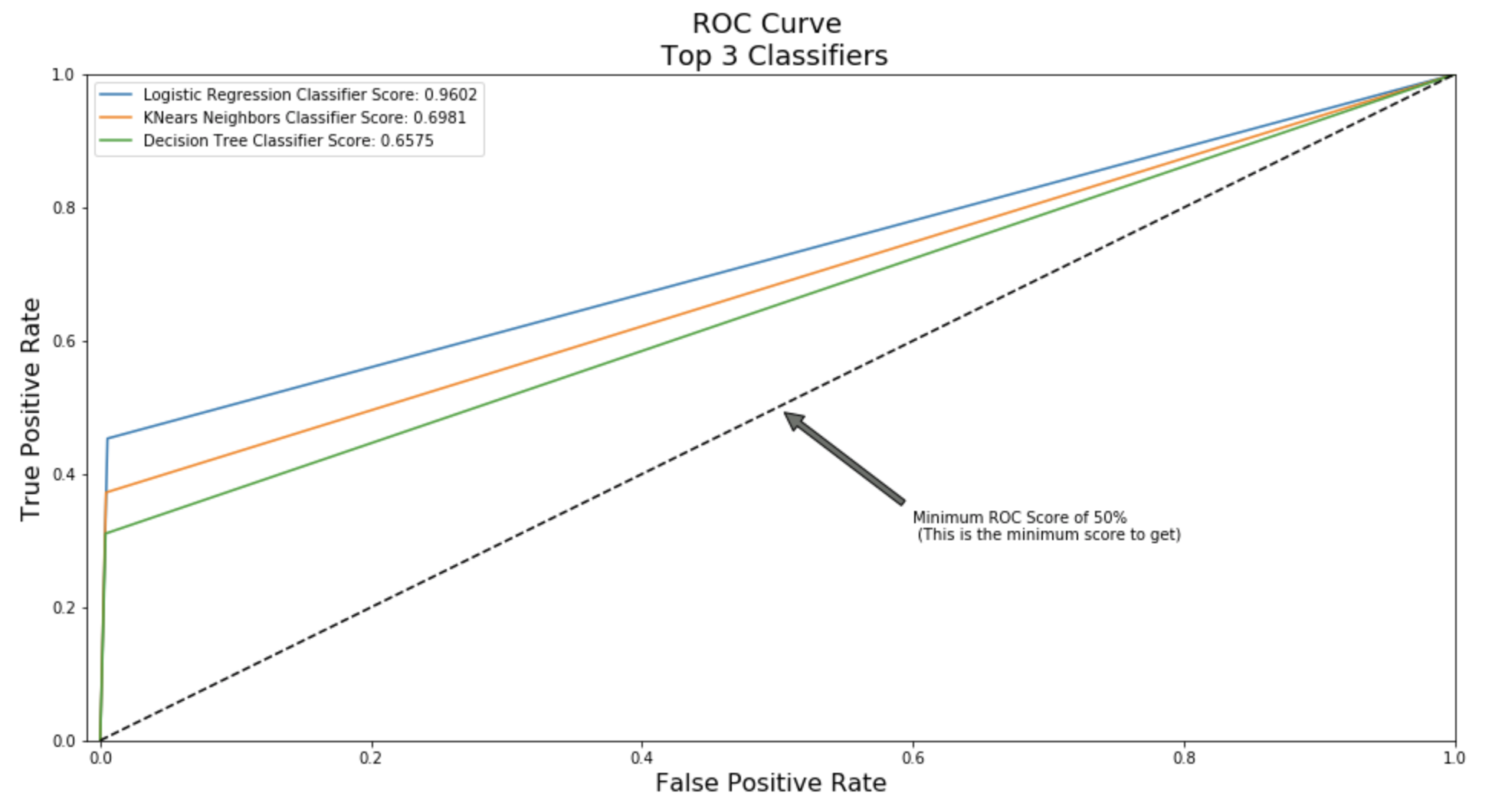






在三种模型的预测能力上，我们能从F1-score 上发现，三种模型对于预测非违约的能力比较稳定，但是对于预测违约的表现上，并不是很突出，这原因来自于原样本数据的不平衡。在三种模型的比较中，我们也能发现logistic regression拥有更稳定的预测能力。

1. **ROC**



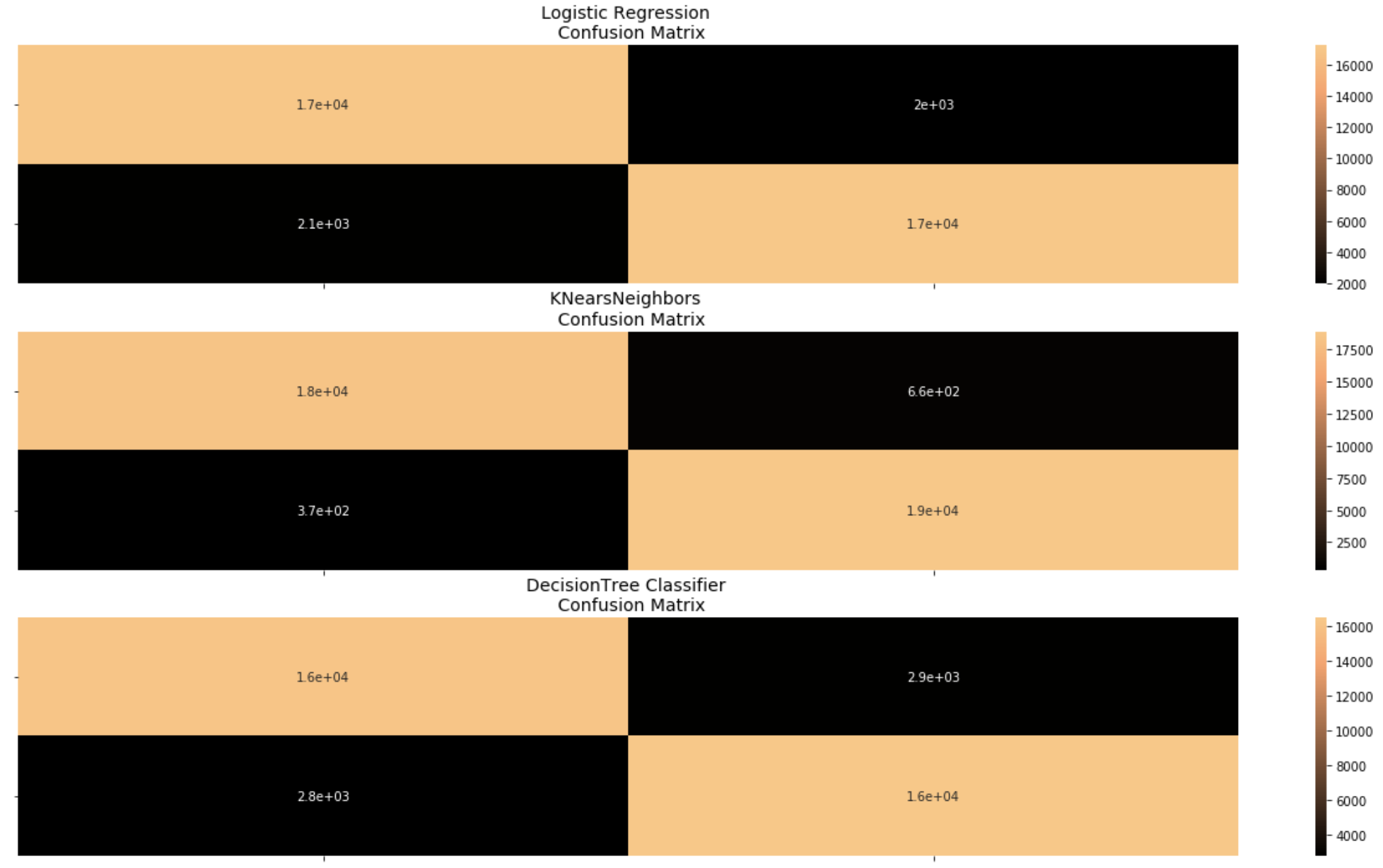
* Logistic 模型拥有最优的预测能力；
* 模型在预测非违约的能力上准确度较高但是对于违规的预测表现不是很稳定；

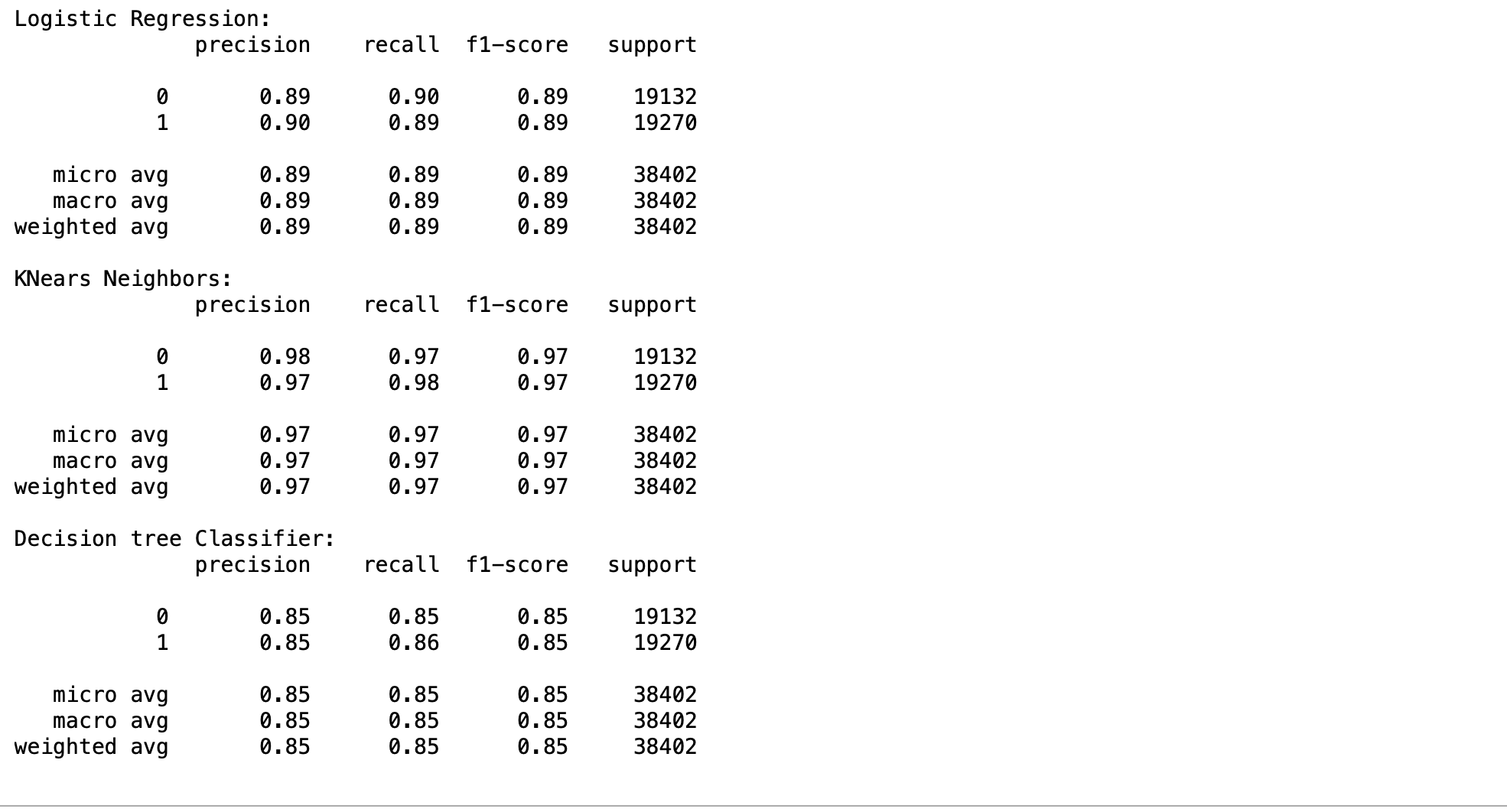
**六、 总结与展望**

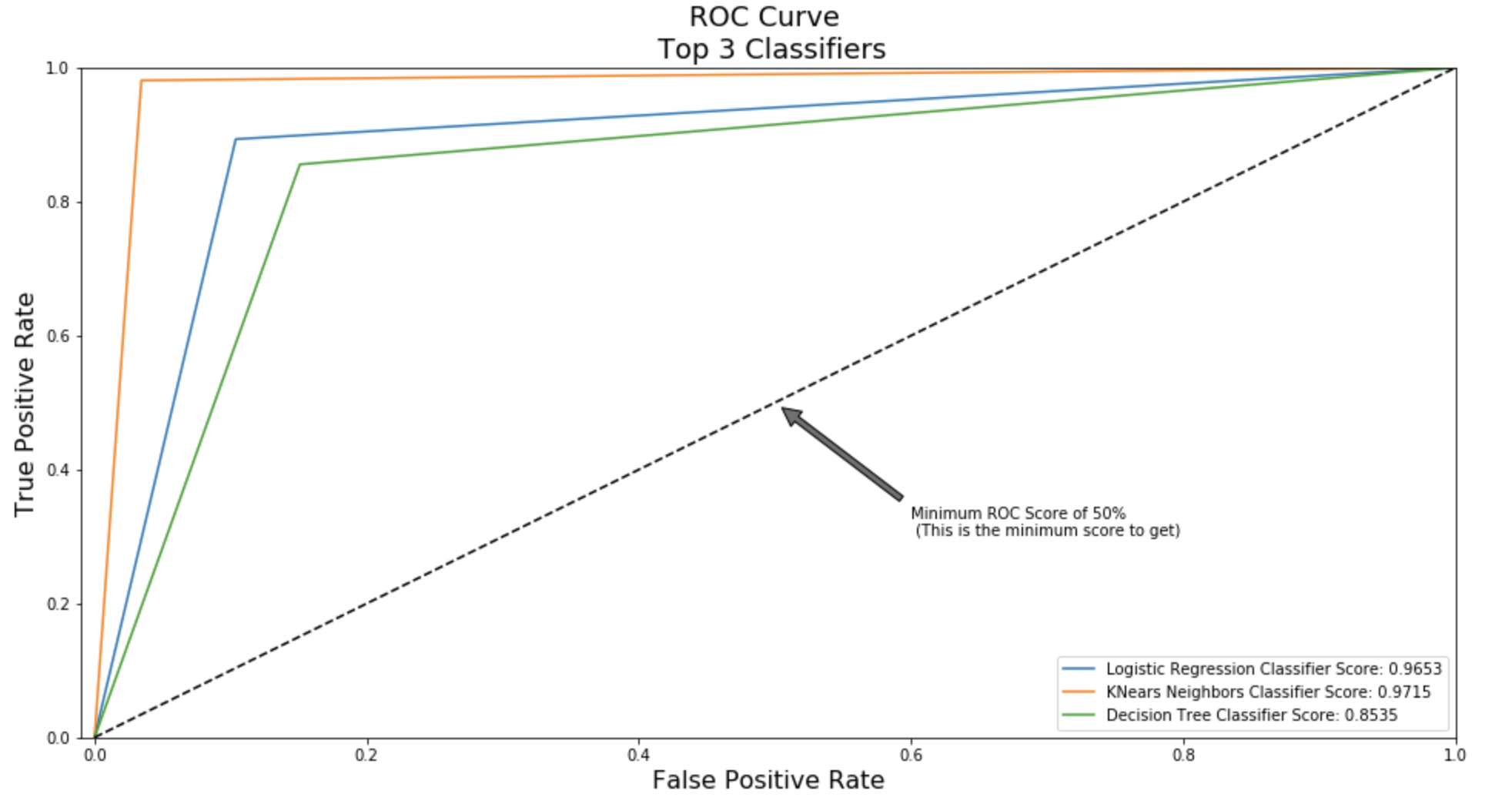
1. **利用平衡的数据集获得的模型**

在这里，我们用平衡的数据集来训练三种模型，下图为在新的训练集下三种

模型的混淆矩阵和ROC 曲线：







我们可以看到，在均衡的样本下，我们三个训练的模型在违约和非违约的预测上都有了较高的的准确性。同时，在roc曲线中，我们能看到knn在三种模型中表现最好。

1. **对特征工程的思考**

* 在本次项目中，考虑到计算能力的有限，我们只选择了15个相关性较大的特

量；

* 特征值中有较多的bool值

1. **总结**

* 本次项目中，logistic模型表现最好，拥有96%的area under auc；
* 本模型在预测非违约（Y = 0）的能力要远高于预测违约的能力（Y = 1），这会对银行带来较大的风险；
* 强烈建议需要对数据集进行重新抽样处理；
* 对样本数据进行SMOTE处理后，得到的模型有更好的预测能力，knn为最优的预测模型；

所有机器学习的完整代码均在压缩文件中，包括没有对样本进行处理的（without\_sampling）和对样本进行处理过的（with\_sampling）,如有任何问题，[可以直接联系我邮箱yanggao131@gmail.com](mailto:可以直接联系我邮箱yanggao131@gmail.com).谢谢！