可信机器学习期末实验报告

实验课程：可信机器学习	年级：2018
实验名称：UCI 信用卡数据集的二元分类分析	姓名：龙旷飞、李泽浩、杨添程
实验编号：期末Project	学号：10172100255、10185102142、10185102204
项目总称：Reducing bias in AI-Based financial Services	指导教师：赵静

一、实验背景

人工智能（AI）为改变我们分配信贷和处理风险的方式提供了一个机会，并创造了更公平、更包容的系统。人工智能可以避免传统的信用报告和评分系统，这有助于抛弃现有的偏见，使它成为一个难得的，改变现状的机会。然而，人工智能很容易朝另一个方向发展，加剧现有的偏见，创造出一个循环，加强有偏见的信贷分配，同时使贷款歧视更难找到。我们将通过开源模型Fairlearn来释放积极的一面，缓解偏见消极的一面。

本次项目中我们模拟了贷款决策中出现的性别准确性差异问题。具体来说，我们考虑的情况是，算法工具根据历史数据进行训练，其对贷款申请人的预测被用于对申请人做出决定。关于涉及信用额度决策中基于性别的歧视的例子，请看这里。

我们使用UCI信用卡数据集。为了这个练习，我们修改了原始的UCI数据集：我们引入了一个合成特征，该特征对女性客户有很强的预测能力，但对男性申请人没有信息。我们拟合了各种预测客户违约的模型。我们表明，一个没有公平意识的训练算法可以导致一个预测器对女性的准确率远远高于男性，而且简单地从训练中删除敏感特征（在这种情况下为性别）是不够的。最后，我们使用fairlearn中的两个缓解算法`ThresholdOptimizer`和`GridSearch`来改进模型。

二、理论知识

1.机器学习的公平性

人工智能和机器学习系统可能会表现出不公平的行为。 定义不公平行为的一种方法是按其损害或对人的影响来定义。 人工智能系统可以造成许多类型的损害。人工智能造成的两种常见损害是：

- 对分配的损害：AI 系统会对特定群体提供或拒绝提供机会、资源或信息。 示例包括招聘、学校招生和贷款。在这些场景中，某个模型在特定人群中挑选优秀候选人的能力可能要强于在其他人群中进行挑选的能力。
- 对服务质量的损害：人工智能系统针对一个人群的工作质量没有针对另一个人群的工作质量好。 例如，语音识别系统针对女性的工作质量可能没有针对男性的工作质量好。

  为了减少人工智能系统中的不公平行为，你必须评估并缓解这些损害。

2.通过Fairlearn进行公平性缓解和评估

Fairlearn 是一个开源 Python 包，它使机器学习系统开发者能够评估其系统的公平性并缓解不公平性。

（注意：公平性是一个社会性的技术难题。 公平性的许多方面（例如公正和正当程序）没有通过量化的公平性指标进行捕获。 另外，许多量化的公平性指标无法同时得到满足。 Fairlearn 开源包的目标是使人类能够评估不同的影响和缓解策略。 最终，由构建人工智能和机器学习模型的人类用户负责根据其应用场景进行权衡。）

Fairlearn 开源包有两个组件：

- 评估仪表板：这是一个 Jupyter 笔记本小组件，用于评估模型的预测如何影响不同的群体。 它还支持使用公平性和性能指标对多个模型进行比较。
- 缓解算法：一组用于在二元分类和回归中缓解不公平性的算法。

借助这些组件，数据科学家和业务主管能够在公平性和性能之间进行权衡，并选择最适合其需求的缓解策略。

3.评估机器学习模型中的公平性

在 Fairlearn 开源包中，公平性通过一种称为“群体公平性”的方法进行概念化，该方法会询问以下问题：哪些群体有遭受损害的风险？ 相关群体（也称为亚群体）是通过 **敏感特征** 或敏感特性定义的。 敏感特征作为名为 `sensitive_features` 的矢量或矩阵传递给 Fairlearn 开源包中的估算器。 此术语的意思是，系统设计者在评估群体公平性时应该对这些特征敏感。

需要注意的是，这些特征是否包含由于私有数据而产生的隐私影响。 但是“敏感”并不意味着这些特征不应被用来进行预测。

（注意：公平性评估并非纯粹的技术练习。 Fairlearn 开源包可帮助你评估模型的公平性，但不会为你执行评估。 Fairlearn 开源包会帮助识别量化指标以评估公平性，但开发人员还必须执行定性分析来评估其自己的模型的公平性。 上面所述的敏感特征是此类定性分析的一个示例。）

在评估阶段，将通过差异指标对公平性进行量化。 **差异指标** 能够以比率或差值的形式评估并比较模型在不同群体中的行为。 Fairlearn 开源包支持两类差异指标：

- 模型性能差异：这些指标集计算所选性能指标的值在不同子群体之间的差异。 示例包括：
  - 准确率差异
  - 错误率差异
  - 精度差异
  - 召回率差异
  - MAE 差异
  - 许多其他差异
- 选择率差异：此指标包含不同子群体之间的选择率差异。 此差异的一个示例是贷款批准率差异。 选择率是指每个分类中归类为 1 的数据点所占的比例（在二元分类中）或者指预测值的分布（在回归中）。

4.减少机器学习模型中的不公平性

Fairlearn 开源包包括了各种不公平性缓解算法。 这些算法支持对预测器行为的一组约束（称为 **奇偶校验约束** 或条件）。 奇偶校验约束要求预测器行为的某些方面在敏感特征所定义的群体（例如不同的种族）之间具有可比性。 Fairlearn 开源包中的缓解算法使用此类奇偶校验约束来缓解所观察到的公平性问题。

（注意：缓解模型中的不公平性意味着降低不公平性，但这种技术上的缓解无法完全消除此不公平性。 Fairlearn 开源包中的不公平性缓解算法可提供建议的缓解策略，以帮助减少机器学习模型中的不公平性，但它们并不是用来完全消除不公平性的解决方案。 每个特定开发人员的机器学习模型可能还有其他应考虑的奇偶校验约束或条件。

使用 Azure 机器学习的开发人员必须自行确定，缓解措施是否充分消除其机器学习模型的预期使用和部署中的任何不公平性。）

Fairlearn 开源包支持下列类型的奇偶校验约束：

奇偶校验约束	目的	机器学习任务
人口统计奇偶校验	缓解分配损害	二元分类、回归
均等几率	诊断分配和服务质量损害	二元分类
均等机会	诊断分配和服务质量损害	二元分类
有界群体损失	缓解服务质量损害	回归

5.缓解算法

（其中本次实验用到了下面的 `GridSearch` 与 `ThresholdOptimizer`方法）
Fairlearn 开源包提供了后期处理和降低不公平性的缓解算法：

算法	说明	机器学习任务	敏感特征
ExponentiatedGradient	公平分类的约简方法中描述的公平分类的黑盒方法	二分类	分类
GridSearch	一种黑盒方法，它通过公平回归：量化的定义和基于约简的算法]中描述的用于有界群体损失的算法实现公平回归的网格搜索变体。	回归	二进制
ThresholdOptimizer	监督式学习中的机会均等性，一文的后期处理算法。 此方法采用现有分类器和敏感特征作为输入，并派生分类器预测的单一转换，以强制实施指定的奇偶校验约束。	二分类	分类

6.ROC&AUC

因为本次项目是一个二分类问题所以涉及到二分类模型的测评指标，下面介绍ROC与AUC指标。ROC（Receiver Operating Characteristic，接受者工作特征曲线）曲线和AUC常被用来评价一个二值分类器（binary classifier）的优劣。

（1）ROC曲线

通过调整模型预测的阈值可以得到不同的点，将这些点可以连成一条曲线，这条曲线叫做接受者工作特征曲线(Receiver Operating Characteristic Curve，简称ROC曲线）。

ROC曲线有一个很好的特征：在实际的数据集中经常会出现类别不平衡现象，即负样本比正样本多很多（或者相反），而且测试数据中的正负样本的分布也可能随着时间而变化。而在这种情况下，ROC曲线能够保持不变。曲线中其中横坐标为FPR（False positive rate 假阳率），纵坐标为真阳率TPR（True postive rate）。**ROC曲线越接近左上角，该分类器的性能越好，**意味着分类器在假阳率很低的同时获得了很高的真阳率。

（2）AUC

AUC（Area Under Curve）被定义为ROC曲线下的面积，显然这个面积的数值不会大于1。又由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围在0.5和1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应AUC更大的分类器效果更好。

AUC是一个概率值，当随机挑选一个正样本以及一个负样本，当前的分类算法根据计算得到的分数将这个正样本排在负样本前面的概率就是AUC值。所以，AUC的值越大，当前的分类算法越有可能将正样本排在负样本值前面，既能够更好的分类。

7.LightGBM

GBDT (Gradient Boosting Decision Tree) 是机器学习中一个长盛不衰的模型，其主要思想是利用弱分类器（决策树）迭代训练以得到最优模型，该模型具有训练效果好、不易过拟合等优点。GBDT不仅在工业界应用广泛，通常被用于多分类、点击率预测、搜索排序等任务；在各种数据挖掘竞赛中也是致命武器，据统计Kaggle上的比赛有一半以上的冠军方案都是基于GBDT。而LightGBM（Light Gradient Boosting Machine）是一个实现GBDT算法的框架，支持高效率的并行训练，并且具有更快的训练速度、更低的内存消耗、更好的准确率、支持分布式可以快速处理海量数据等优点。

三、实验内容概述

- **涉及领域：**
  - 金融贷款方面的决策分析。我们分析的数据是原始数据经过人工简单处理过的，是为了展现准确性方面的悬殊差异。
- **机器学习任务：**
  - 二元性分类
- **机器学习公平任务：**
  - 使用Fairlearn metrics和Fairlearn dashboard来评估模型的公平。
  - 使用Fairlearn中的改进算法来改进模型的公平水平。
- **性能指标：**
  - ROC曲线下的面积。
  - 平衡过后的准确率。
- **公平指标：**
  - Equalized-odds difference.
- **改进的算法:**
  - `fairlearn.reductions.GridSearch`
  - `fairlearn.postprocessing.ThresholdOptimizer`

四、数据分析

UCI数据集包含30,000名客户及其在台湾一家银行的信用卡交易数据。除了客户的静态特征外，该数据集还包含某年4月至9月的信用卡账单支付历史，以及客户信用卡的余额限制。目标是客户是否会在接下来的一个月，即该年10月拖欠信用卡付款。可以想象，在这个数据上训练出来的模型在实践中可以用来确定客户是否有资格获得其他产品，如汽车贷款等。

该数据集包含23个输入变量(input variable)和一个响应变量（response variable）。该数据集来源于UCI machine learning repository,为某银行的信用卡客户信息数据，共有30000个样本，包括过去六个月的账单还款情况。

ID：信用卡客户ID号
LIMIT_BAL：以新台币计算的信贷金额（包括个人和家庭/补充信贷）/ 信用卡限额，会被替换成一个合成的更具典型性的特征。
SEX： 性别 (1代表男性，2代表女性)
EDUCATION：受 教育程度(1=研究生, 2=大学, 3=高中, 4=其他 5=未知, 6=未知)
MARRIAGE：婚姻状况（1=已婚，2=单身，3=其他）
AGE：年龄
X1：信用额度，包括其个人和家庭补充信用
X2：性别（1=male;2=female）
X3：教育（1=研究生，2=大学，3=高中，4=其他）
X4：婚姻状况（1=已婚，2=单身，3=其他）
X5：年龄，age
X6-X11：过去六个月的还款情况。X6-X11为9-4月的还款情况。其中，-1,代表按时还款；1,代表延时一个月还款；2,代表延时两个月还款.....依次类推，XN=n,代表延时n个月还款，
X12-X17：过去六个月的账单数额情况。X12-X17为9-4月账单数额情况
X18-X23：过去六个月的还款数额情况。X18-X23为9-4月还款数额情况
Y：目标属性，客户下个月还款违约情况（1=逾期，0=未逾期）

实践部分：
read_csv读入数据集后
（1）使用 ‘data.info()’ 查看数据整体信息

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  float64
 2   SEX                         30000 non-null  int64
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  int64
 5   AGE                         30000 non-null  int64
 6   PAY_0                       30000 non-null  int64
 7   PAY_2                       30000 non-null  int64
 8   PAY_3                       30000 non-null  int64
 9   PAY_4                       30000 non-null  int64
 10  PAY_5                       30000 non-null  int64
 11  PAY_6                       30000 non-null  int64
 12  BILL_AMT1                   30000 non-null  float64
 13  BILL_AMT2                   30000 non-null  float64
 14  BILL_AMT3                   30000 non-null  float64
 15  BILL_AMT4                   30000 non-null  float64
 16  BILL_AMT5                   30000 non-null  float64
 17  BILL_AMT6                   30000 non-null  float64
 18  PAY_AMT1                    30000 non-null  float64
 19  PAY_AMT2                    30000 non-null  float64
 20  PAY_AMT3                    30000 non-null  float64
 21  PAY_AMT4                    30000 non-null  float64
 22  PAY_AMT5                    30000 non-null  float64
 23  PAY_AMT6                    30000 non-null  float64
 24  default.payment.next.month  30000 non-null  int64
dtypes: float64(13), int64(12)
```

（2）使用 'data.describe()' 查看数据基本情况(包括中位数、最大最小值等)，在此只列出头尾部分

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	3
count	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	30000.000000	
mean	15000.500000	167484.322667	1.603733	1.853133	1.551867	35.485500	-0.016700	
std	8660.398374	129747.661567	0.489129	0.790349	0.521970	9.217904	1.123802	
min	1.000000	10000.000000	1.000000	0.000000	0.000000	21.000000	-2.000000	
25%	7500.750000	50000.000000	1.000000	1.000000	1.000000	28.000000	-1.000000	
50%	15000.500000	140000.000000	2.000000	2.000000	2.000000	34.000000	0.000000	
75%	22500.250000	240000.000000	2.000000	2.000000	2.000000	41.000000	0.000000	
max	30000.000000	1000000.000000	2.000000	6.000000	3.000000	79.000000	8.000000	

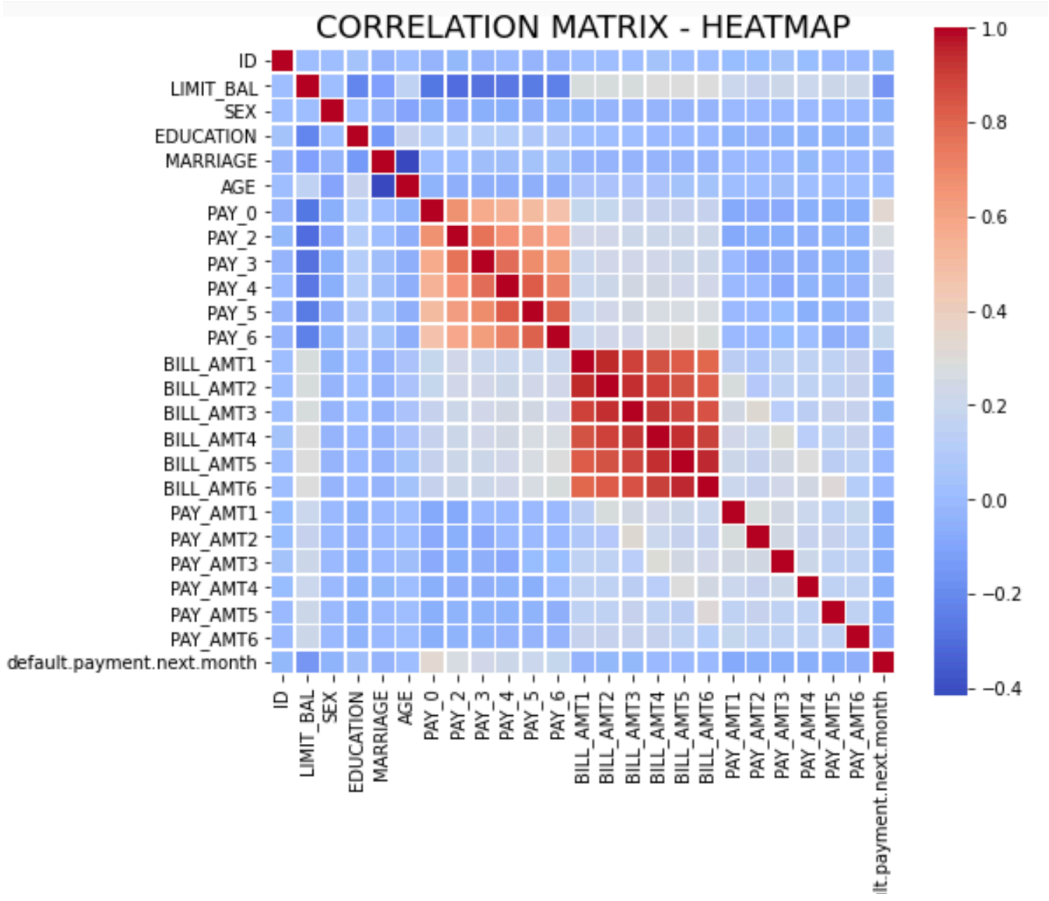
AY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default.payment.next.month
0.000000	3.000000e+04	30000.00000	30000.000000	30000.000000	30000.000000	30000.000000
3.580500	5.921163e+03	5225.68150	4826.076867	4799.387633	5215.502567	0.221200
3.280354	2.304087e+04	17606.96147	15666.159744	15278.305679	17777.465775	0.415062
0.000000	0.000000e+00	0.00000	0.000000	0.000000	0.000000	0.000000
0.000000	8.330000e+02	390.00000	296.000000	252.500000	117.750000	0.000000
0.000000	2.009000e+03	1800.00000	1500.000000	1500.000000	1500.000000	0.000000
6.000000	5.000000e+03	4505.00000	4013.250000	4031.500000	4000.000000	0.000000
2.000000	1.684259e+06	896040.00000	621000.000000	426529.000000	528666.000000	1.000000

（3）使用 'data.isnull().sum()' 查看每个属性是否存在缺失值

```
ID                                  0
LIMIT_BAL                           0
SEX                                 0
EDUCATION                           0
MARRIAGE                            0
AGE                                 0
PAY_0                               0
PAY_2                               0
PAY_3                               0
PAY_4                               0
PAY_5                               0
PAY_6                               0
BILL_AMT1                           0
BILL_AMT2                           0
BILL_AMT3                           0
BILL_AMT4                           0
BILL_AMT5                           0
BILL_AMT6                           0
PAY_AMT1                            0
PAY_AMT2                            0
PAY_AMT3                            0
PAY_AMT4                            0
PAY_AMT5                            0
PAY_AMT6                            0
default.payment.next.month          0
dtype: int64
```

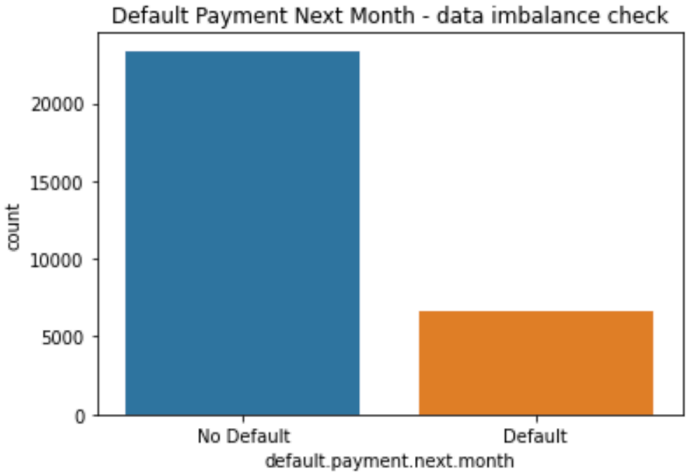
（4）为了使属性对于预测结果的影响更加直观，下面进行数据可视化，对属性进行分析

使用皮尔逊函数定义的相关矩阵观察所有属性与预测结果的相关性，颜色越深代表相关度越高

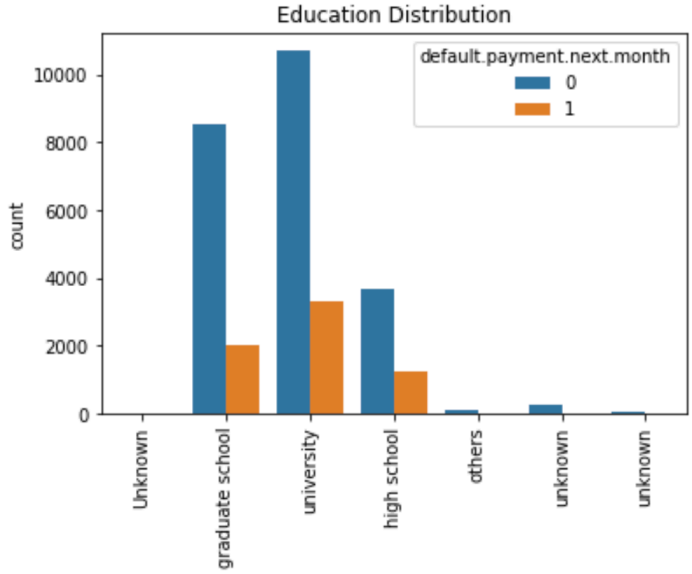


可以看到PAY_0 - PAY_6与违约情况相关度最高

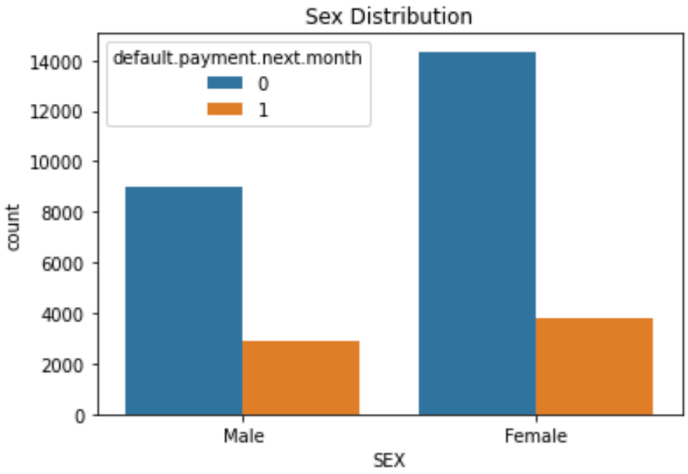
总体查看下个月违约情况



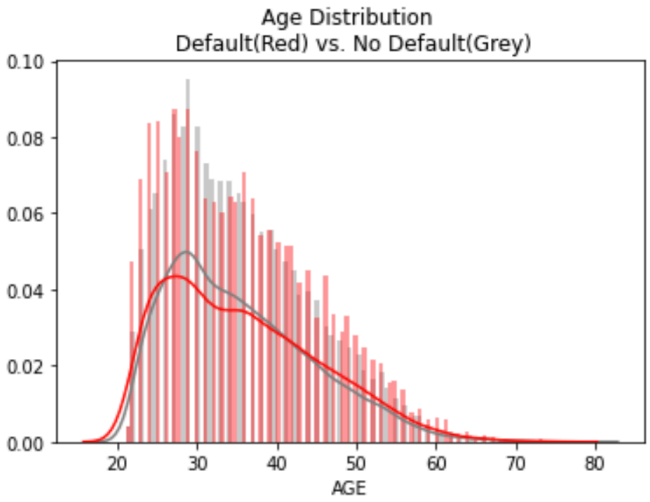
受教育水平分布与违约情况



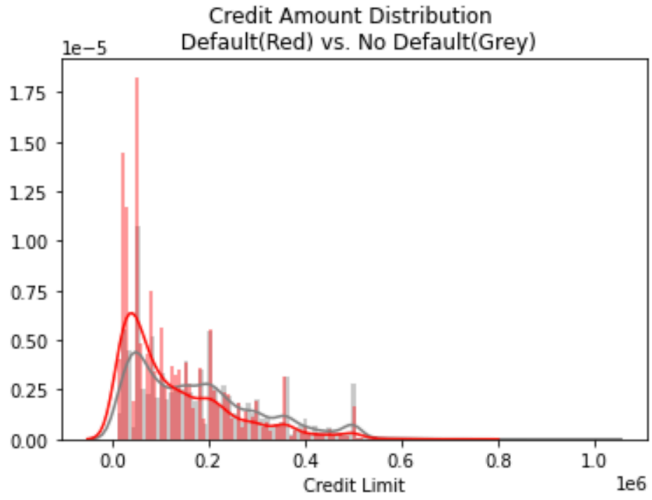
性别分布与违约情况



年龄分布与与违约情况



信用额度分配



五、实验过程

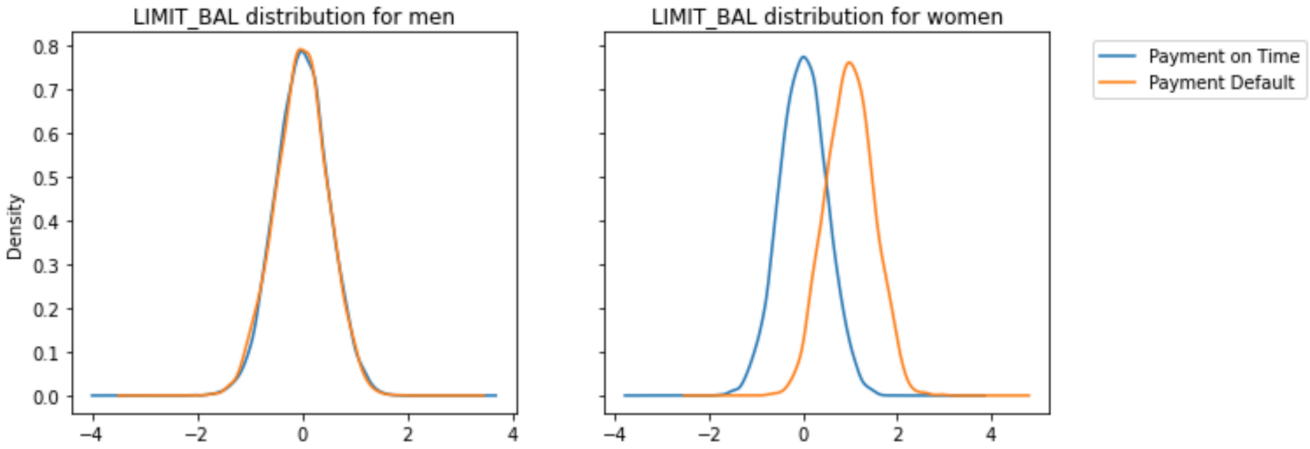
1.修改特征

我们魔改一下信用卡限额信息 `LIMIT_BAL` 使得该特征对于预测女人来说非常“强”而对男人来说相对“弱”。例如，我们可以想象，较高的信用额度表明女性客户违约的可能性较小，但对男性客户的违约概率没有提供任何信息。

```
dist_scale = 0.5
np.random.seed(12345)
# Make 'LIMIT_BAL' informative of the target
dataset['LIMIT_BAL'] = Y + np.random.normal(scale=dist_scale, size=dataset.shape[0])
# But then make it uninformative for the male clients
dataset.loc[A==1, 'LIMIT_BAL'] = np.random.normal(scale=dist_scale,
size=dataset[A==1].shape[0])

fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(10, 4), sharey=True)
# Plot distribution of LIMIT_BAL for men
dataset['LIMIT_BAL'][(A==1) & (Y==0)].plot(kind='kde', label="Payment on Time", ax=ax1,
```

```
                                                title="LIMIT_BAL distribution for men")
dataset['LIMIT_BAL'][(A==1) & (Y==1)].plot(kind='kde', label="Payment Default", ax=ax1)
# Plot distribution of LIMIT_BAL for women
dataset['LIMIT_BAL'][(A==2) & (Y==0)].plot(kind='kde', label="Payment on Time", ax=ax2,
                                                legend=True, title="LIMIT_BAL distribution
for women")
dataset['LIMIT_BAL'][(A==2) & (Y==1)].plot(kind='kde', label="Payment Default", ax=ax2,
                                                legend=True).legend(bbox_to_anchor=(1.6, 1))
plt.show()
```



我们从上面的数字和图像中注意到，新的 'LIMIT_BAL' 功能对女性确实有很高的预测性，但对男性则没有。

2.使用无公平意识的模型

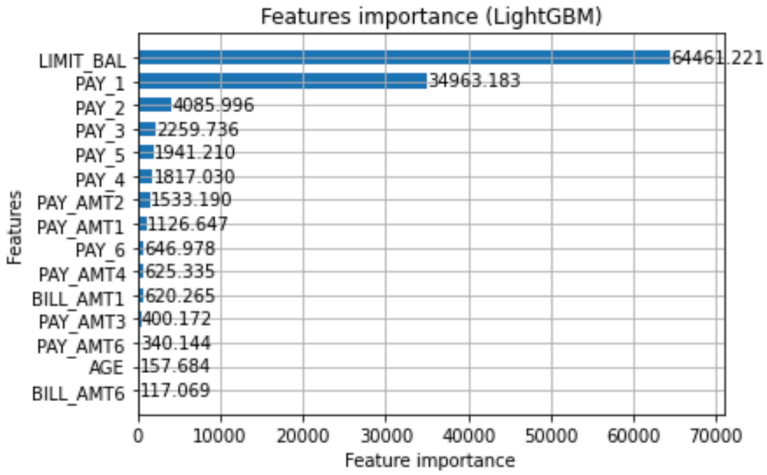
我们在修改后的数据上训练一个开箱即用的 'lightgbm' 模型，并评估几个差异指标。

计算AUC值

```
roc_auc_score(Y_train, model.predict_proba(df_train)[:, 1])
```

```
roc_auc_score(Y_trai
Out[11]: 0.8500200312035275
```

可视化模型中各特征值对重要程度，可以看到合成特征 'LIMIT_BAL' 在这个模型中作为最重要的特征出现，尽管它对数据中的整个人口段没有预测能力。



我们注意到人工修改过的特征 `LIMIT_BAL` 作为这个模型中最重要的特征出现，尽管它对数据中的整个人口部分没有预测能力。

计算几个性能和差异指标：

	Unmitigated
Overall selection rate	0.267111
Demographic parity difference	0.0499407
Demographic parity ratio	0.825666

Overall balanced error rate	0.22133
Balanced error rate difference	0.178303
Equalized odds difference	0.344951

Overall AUC	0.852206
AUC difference	0.189891

从上到下依次为：总体选择率、人口统计学上的均等差异、人口统计学上的奇偶比、总体平衡错误率、平衡误差率差、均衡赔率差、总体AUC、AUC差异

作为总体性能指标，我们使用ROC曲线下的面积（AUC），它适合于正负例子之间有很大不平衡的分类问题。对于二元分类器来说，这与平衡精度是一样的。

作为公平性的衡量标准，我们使用*equalized odds difference*，它量化了不同人口统计学中所经历的准确性的差异。我们的目标是确保两组中的任何一组（男性对女性）的假阳性率或假阴性率都不比另一组大。*equalized odds difference*等于以下两个数字中较大的一个。1）两组的假阳性率之差；（2）两组的假阴性率之差。

上表显示总体AUC为0.85（基于连续预测），总体平衡误差率(overall balanced error)为0.22（基于0/1预测）。在我们的应用环境中，这两点都是令人满意的。然而，准确率有很大的差距,如平衡误差率差异(balanced error rate difference)，当我们考虑到均衡赔率差异(equalized-odds difference)时，差距甚至更大。作为理智的检查，我们还显示了人口学的平准率，其水平（略高于0.8）在这种情况下被认为是令人满意的。

3.通过后处理减小Equalized-Odds Difference

我们尝试缓解 `lightgbm` 预测中的基于敏感特征的结果悬殊，方法是使用Fairlearn postprocessing算法 `ThresholdOptimizer`，该算法通过在equalized-odds difference（在训练数据上）为零的约束条件下优化准确率，为 `lightgbm` 模型产生的分数（类别概率）找到一个合适的阈值。由于我们的目标是优化均衡准确率，我们对训练数据重新取样，使其具有相同数量的正面和负面例子。这意味着 `ThresholdOptimizer` 对于优化在原数据上取得的平衡准确率是非常有效的。

```
postprocess_est = ThresholdOptimizer(estimator=model, constraints="equalized_odds")

balanced_idx1 = df_train[Y_train==1].index
pp_train_idx = balanced_idx1.union(Y_train[Y_train==0].sample(n=balanced_idx1.size,
random_state=1234).index)

df_train_balanced = df_train.loc[pp_train_idx, :]
Y_train_balanced = Y_train.loc[pp_train_idx]
A_train_balanced = A_train.loc[pp_train_idx]

postprocess_est.fit(df_train_balanced,Y_train_balanced,sensitive_features=A_train_balan
ced)
```

`ThresholdOptimizer`算法大大减少了原有的不均衡性。然而，性能指标（平衡错误率以及AUC）变得更糟。在实践中部署这样一个模型之前，重要的是要更详细地研究为什么我们观察到这样的情况。在我们的案例中，这是因为可用的特征对其中一个人口群体的信息量比对另一个人口群体的信息量小得多。

跟不管公平的算法相比`ThresholdOptimizer`产生0/1预测, 所以它的平衡误差率之差等于AUC之差，而它的总体平衡误差率等于1-总体AUC。

接下来，我们在dashboard中比较`lightgbm`模型和这个模型。作为性能指标，我们可以选择均衡准确率。现在的仪表板并不直接显示均衡的赔率差异，但在准确率差异视图中显示了类似的信息，我们可以检查两组的高预测率和低预测率之间的差异。

4.用改进 GridSearch 算法改进 Equalized-Odds Difference

我们现在尝试使用 `GridSearch` 算法来缓解差异。与 `ThresholdOptimizer` 不同， `GridSearch` 产生的预测器在测试时不访问敏感特征。另外，我们不是训练单一的模型，而是训练与性能指标（平衡精度）和公平指标（均衡赔率差异）之间的不同权衡点相对应的多个模型。

```
sweep = GridSearch(model,
                   constraints=EqualizedOdds(),
                   grid_size=41,
                   grid_limit=2)

sweep.fit(df_train_balanced, Y_train_balanced, sensitive_features=A_train_balanced)

sweep_preds = [predictor.predict(df_test) for predictor in sweep.predictors_]
sweep_scores = [predictor.predict_proba(df_test)[:, 1] for predictor in
sweep.predictors_]

equalized_odds_sweep = [
```

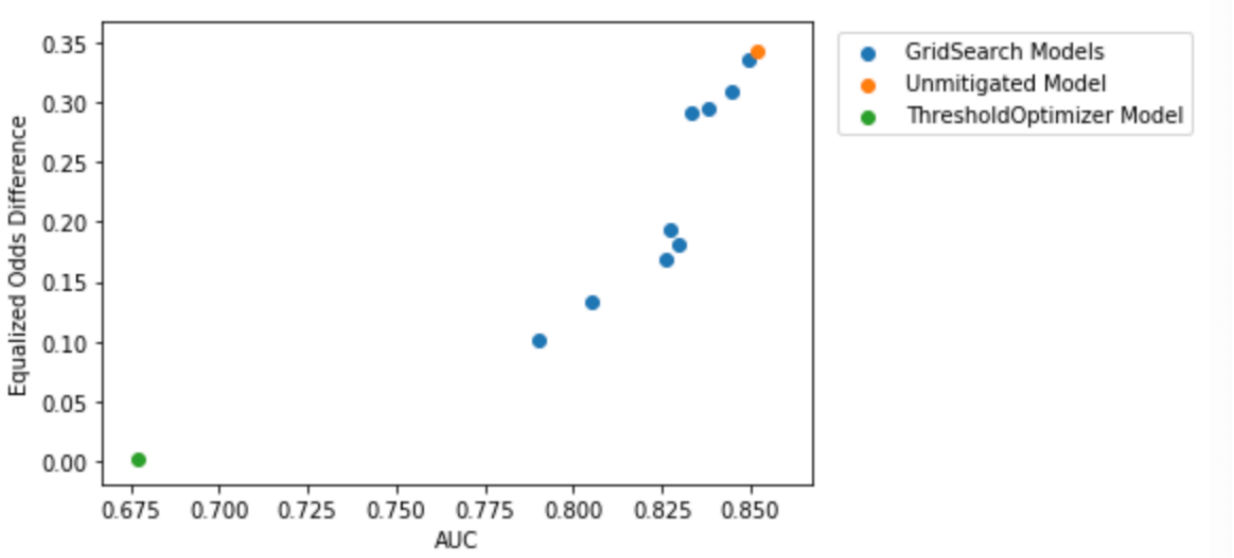
```
    equalized_odds_difference(Y_test, preds, sensitive_features=A_str_test)
    for preds in sweep_preds
]
balanced_accuracy_sweep = [balanced_accuracy_score(Y_test, preds) for preds in
sweep_preds]
auc_sweep = [roc_auc_score(Y_test, scores) for scores in sweep_scores]
```

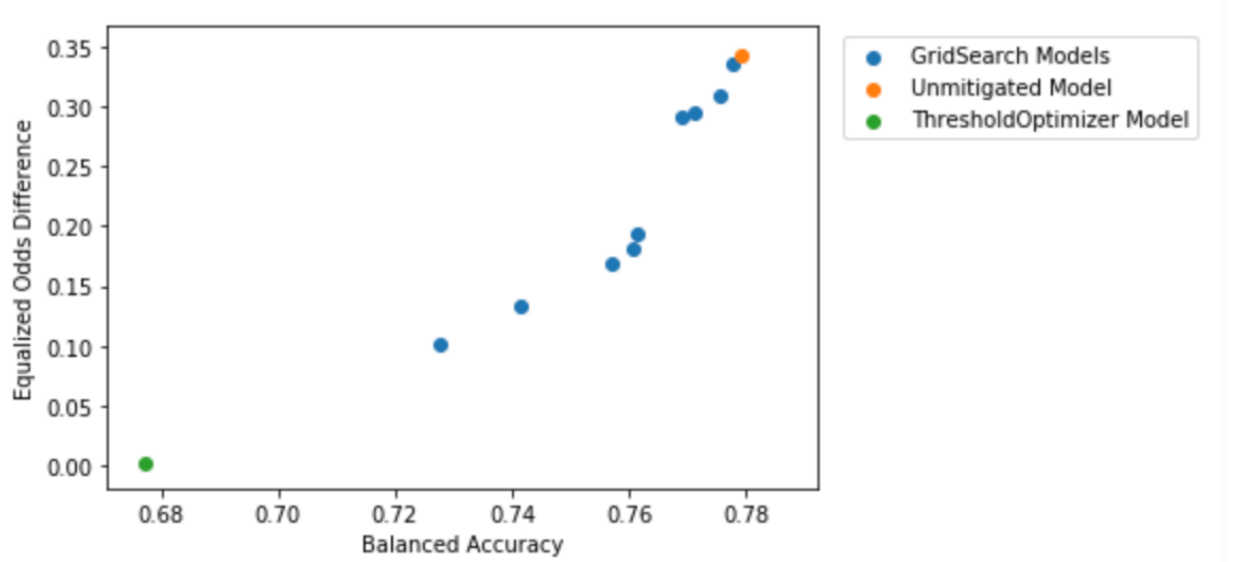
5.结果分析

Balanced Accuracy & Equalized-Odds Difference

正如预期的那样，`GridSearch` 模型沿着增大的均衡准确率（但也存在更大的差异问题）如下。这使数据科学家可以灵活地选择一个最适合应用环境的模型，取得均衡准确率与差异中的一个可以接受的平衡点。

同样我们在AUC与差异之间也可以选择一个可以接受的平衡点。





在Fairlearn提供的Dashboard中可以看到每个模型的accuracy/Precision等细节，能帮助我们选择适合的模型。

六、总结

实现了一个从公平入手的简单的可信机器学习算法。对于我们小组成员受益匪浅，但是人工智能要想具备真正而完备的公平性，还有很长的路要走。

在2019年底，美国联邦地方法院裁定，哈佛大学在针对亚裔申请者的歧视案中获得胜诉。该项起诉的结果宣告，哈佛大学故意压低亚裔美籍学生的录取数量、涉嫌种族歧视的行为，在政府层面获得了支持。原本是有利于保护少数族裔机会均等和族裔平衡的政策，却在一定程度上损害了亚裔族群的权益。可见公平性领域还有许多未解难题：比如怎样兼顾平等和公平？有没有一种公平性准则可以更好地照顾各方利益?

算法的公平性从本质上来讲是一个宏大的命题，远远不止包含数据和算法。公平性问题的层级在数据和模型之上，还有公平性的定义、道德伦理/法律。

算法公平性的落实，需要政府监管方、行业专家、科研开发者、用户的共同努力。首先，政府监管方与行业专家，根据行业需求制定合理的公平性准则，并制定算法歧视的问责法律。然后，科研开发者在此基础上通过设计实现算法公平。在算法设计之初，就将算法公平性准则、算法可解释性、算法问责等价值需求囊括在算法设计之中，这也督促设计者在设计阶段严格遵守公平性的伦理和法律规则。最后，在算法的应用阶段，政府监管方与用户，共同监督算法公平性实施。对严重的算法歧视行为，由政府监管方进行问责。通过各方联合起来，解决算法的不公平问题，人工智能才能够被放心地应用于民生中的各个领域，并真正地造福全社会。

七、附录

【1】https://docs.microsoft.com/zh-cn/azure/machine-learning/concept-fairness-ml

【2】https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/