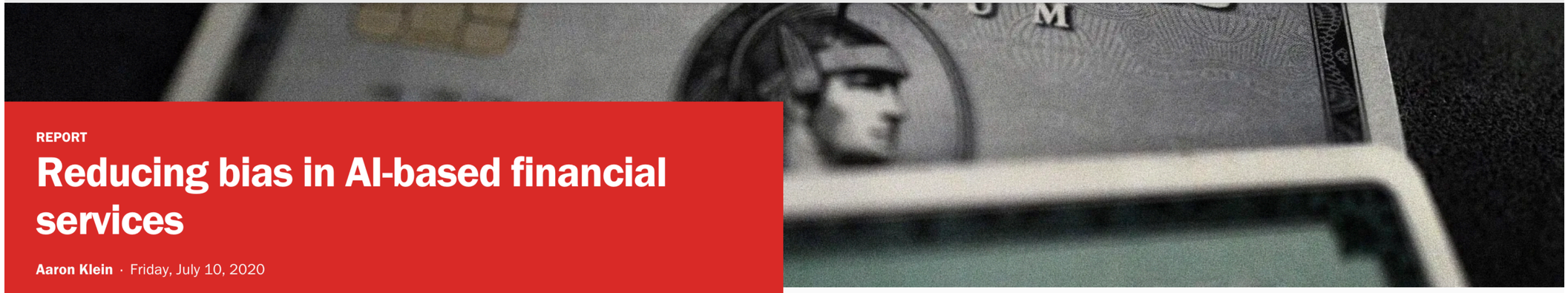
可信机器学习期末作业

基于信贷数据的公平学习模型

龙旷飞/李泽浩/杨添程



REPORT

Reducing bias in AI-based financial services

Aaron Klein · Friday, July 10, 2020

For media inquiries, contact:

Governance Studies Main Line
202.797.6090

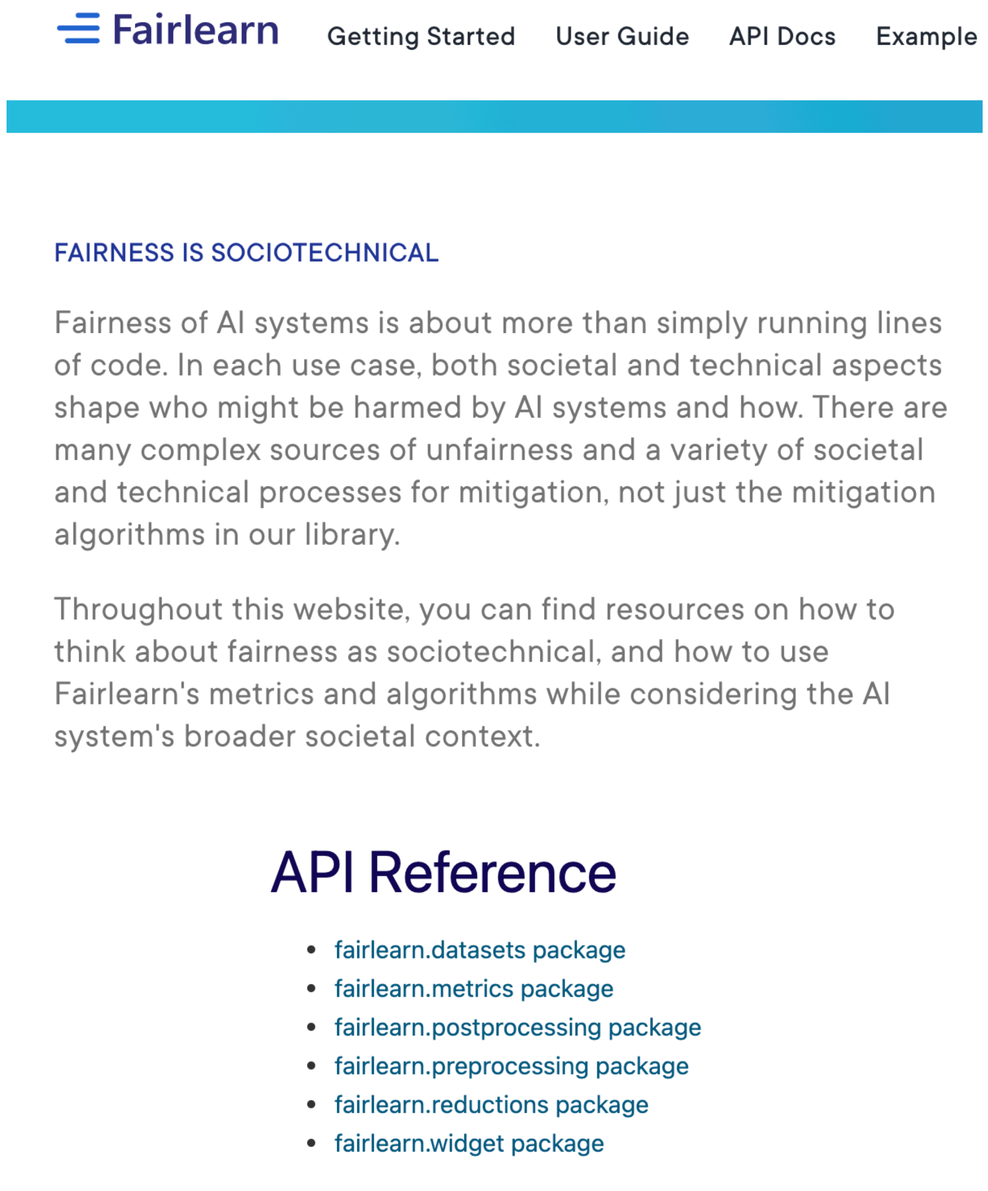
Editor's Note: *This report from The Brookings Institution's Artificial Intelligence and Emerging Technology (AIET) Initiative is part of "AI Governance," a series that identifies key governance and norm issues related to AI and proposes policy remedies to address the complex challenges associated with emerging technologies.*

Artificial intelligence (AI) presents an opportunity to transform how we allocate credit and risk, and to create fairer, more inclusive systems. AI's ability to avoid the traditional credit reporting and scoring system that helps perpetuate existing bias makes it a rare, if not unique, opportunity to alter the status quo. However, AI can easily go in the other direction to exacerbate existing bias, creating cycles that reinforce biased credit allocation while making discrimination in lending even harder to find. Will we unlock the positive, worsen the negative, or maintain the status quo by embracing new technology?

基于信贷数据的公平学习模型

Fairlearn库

- Fairlearn是一个开源的、社区驱动的项目，帮助数据科学家提高人工智能系统的公平性。
- 在整个开源库中，你可以找到关于如何将公平性视为社会技术的资源，以及如何在考虑人工智能系统更广泛的社会背景下使用Fairlearn的指标和算法。



基于信贷数据的公平学习模型

评估公平性

- 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。

- 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：模型性能差异和选择率差异。

基于信贷数据的公平学习模型

减轻模型中的不公平性

- 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。

- 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。

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

基于信贷数据的公平学习模型

实验简介

- 本次项目中我们模拟了贷款决策中出现的准确性差异问题。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。

- 我们使用UCI数据卡数据集。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。

基于信贷数据的公平学习模型

数据分析

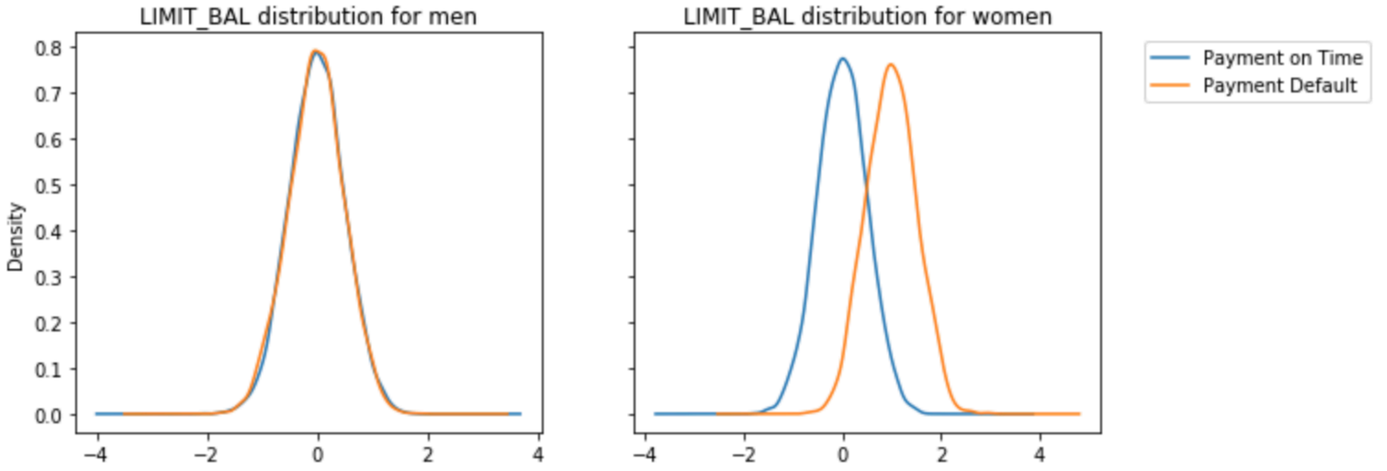
基于信贷数据的公平学习模型

特征修改

- 我们修改一下信用卡的限额信息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。

- 我们从图像中注意到，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。

```
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])
```



基于信贷数据的公平学习模型

无公平意识的模型LightGBM

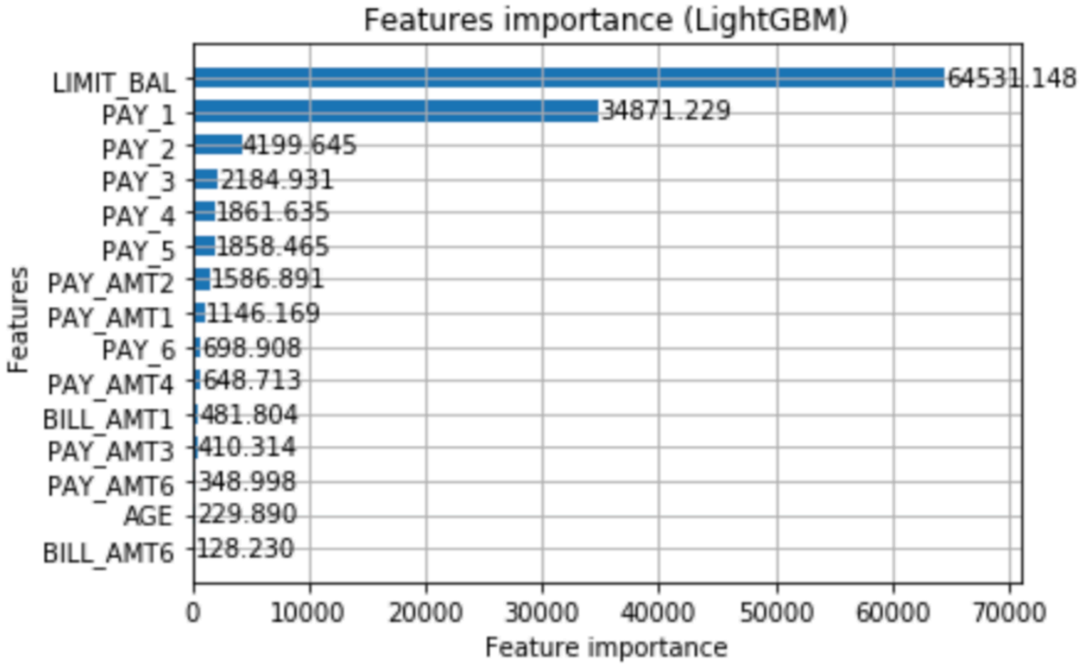
- 从上到下依次为：总体选择率、人口统计学上的均等差异、人口统计学上的奇偶比、总体平衡错误率、平衡误差率差、均衡赔率差、总体AUC、AUC差异。

- 注意人工修改过特征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。

	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



基于信贷数据的公平学习模型

PostProcessing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。

- 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。

基于信贷数据的公平学习模型

GridSearch算法改进Equalized-Odds Difference

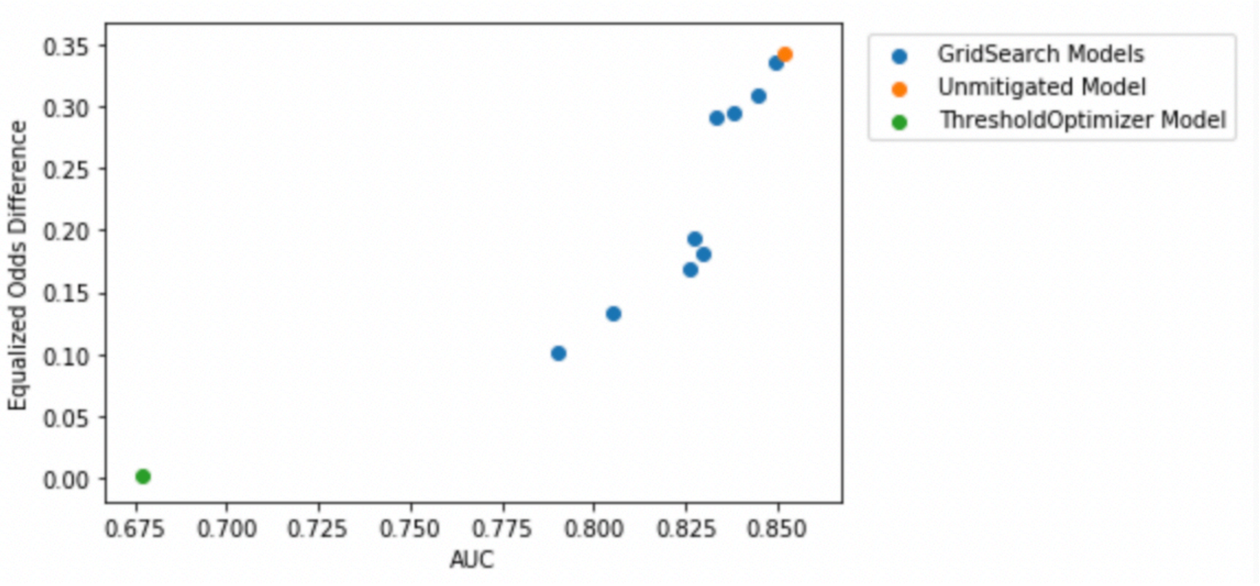
- 我们现在尝试使用GridSearch算法来缓解差异。与ThresholdOptimizer不同，GridSearch产生的预测器在测试时不访问敏感特征。另外，我们不是训练单一的模型，而是训练与性能指标（平衡精度）和公平指标（均衡赔率差异）之间的不同权衡点相对应的多个模型。

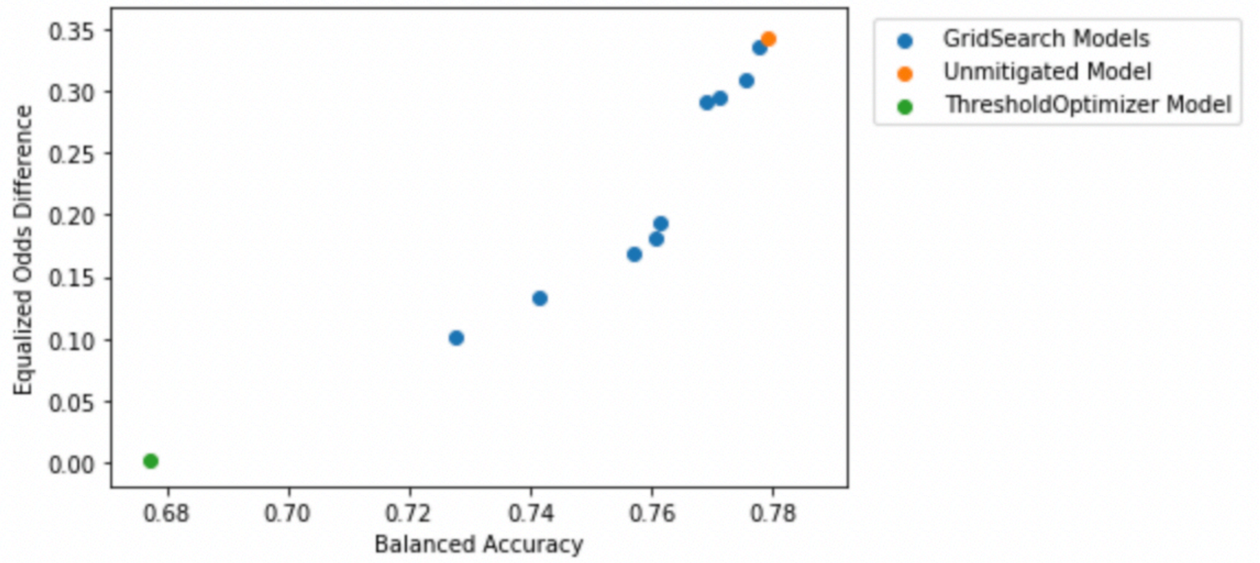
基于信贷数据的公平学习模型

平衡差异与准确率

- 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。

- 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。





谢谢