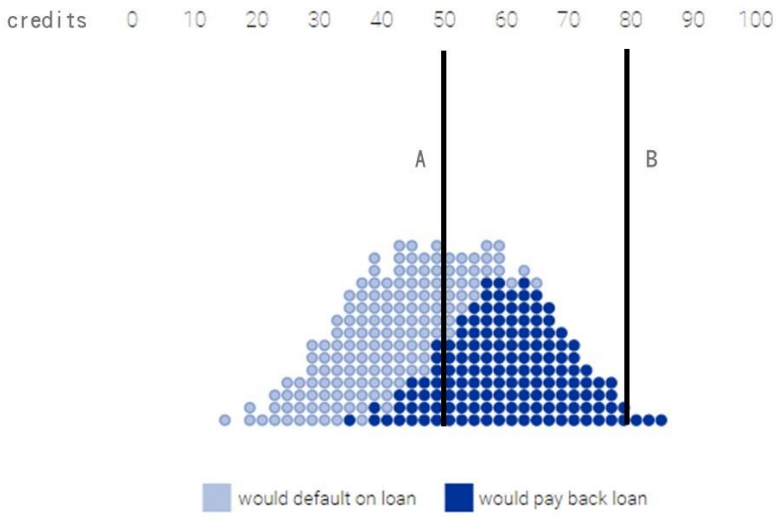
標題：機器之前人人平等？

關鍵字：決策樹墩、演算法公平性、機會均等、What-If Tool

機器學習的演算法一定公平嗎？會不會因為大量利用人類的歷史資料進行訓練，導致無形間沿襲了人類社會中難免存在的歧視與偏見？又會不會因為少數族群提供的訓練資料有限，造成機器學習的判斷不夠準確，因此犧牲他們應該擁有的權利？假設這些問題不幸發生了，再加上機器的自動化，將很有可能加速惡化社會上的種種不公，影響的程度恐怕難以估量，這也就是為什麼「演算法的公平性」逐漸受到重視。以下將以「決策樹墩」為例，簡單說明我們該如何維持分類系統的公平性。

決策樹墩

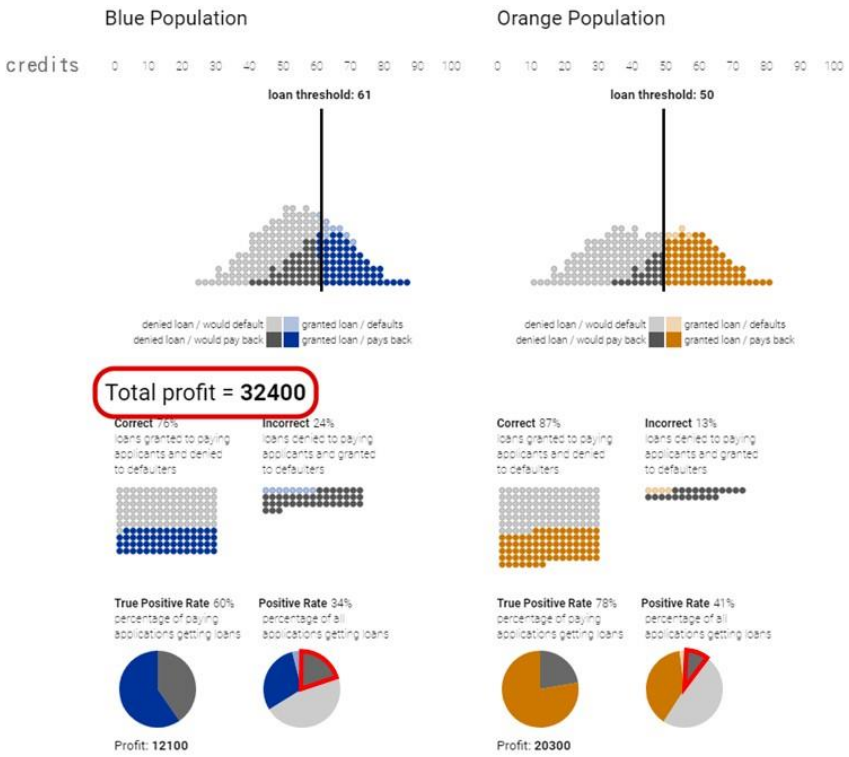
決策樹墩(decision stump，亦稱作 threshold classifier)是一種簡易的二元分類法，通常會根據某單一特徵將資料分成兩堆。舉例來說，我們假設有個利用決策樹墩來核准「租貸資格」的系統，並以貸款者過去的「信用指數」作為分類的依據，唯有信用高於閾值(threshold)的人，方可獲得貸款。



(圖一、單一族群的信用統計圖表，深藍表示會還款，淺藍表示不會還款)

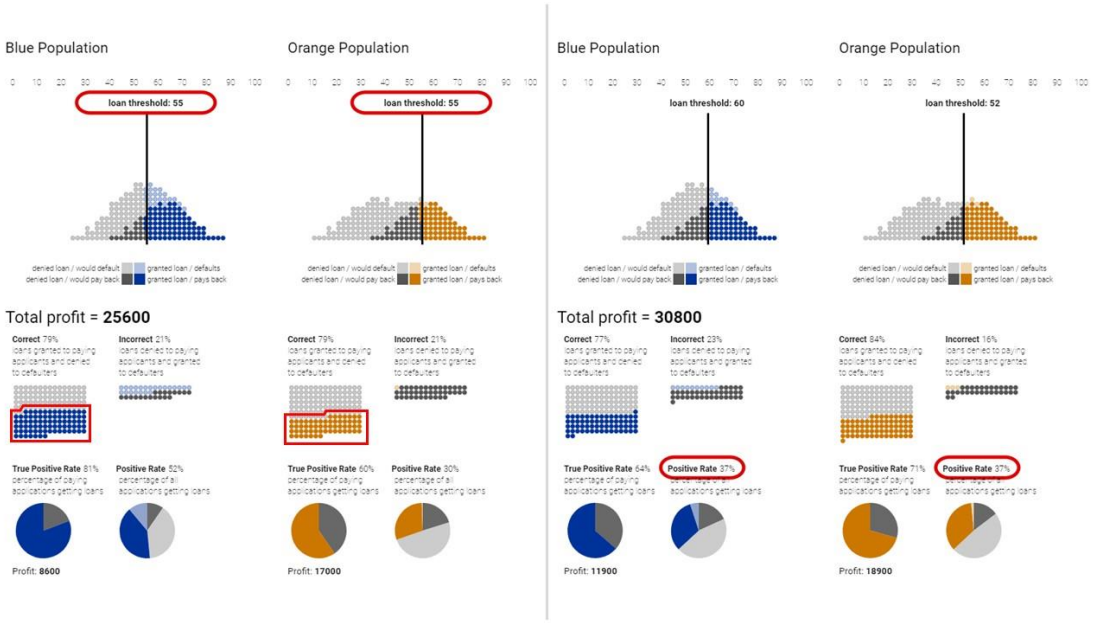
然而，用這些分類依據始終沒有辦法詮釋極為複雜的現實社會，貸款人信用高並不代表保證會還款，信用低的人事實上也不一定會拖債(圖一)。另外，設定不同的閾值也會影響分類的正確性和淨利的多寡，太高的閾值會使公司獲得的淨利過低(圖一中 B 閾值只能賺三個人)，而標準太低又會增加評判的失誤率(圖一中 A 閾值右方的淺藍色點比 B 多)。因此，選擇一個合適的閾值相當重要，非但可以滿足更多守信的客戶，同時也增長租貸公司的盈利能力。

公平問題



(圖二、效益最大化的分析圖表，深彩色表示具有貸款資格而且會還款，淺藍色代表具有貸款資格但不會還款，深灰色代表不具有貸款資格但本來會還款，淺灰色代表不具有貸款資格而且本來不會還款)

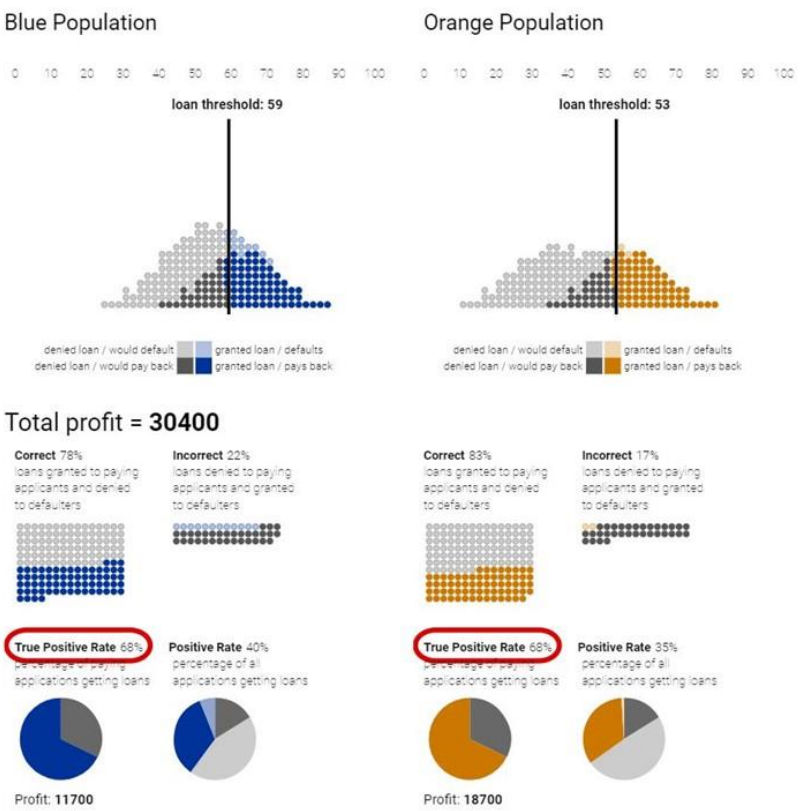
找到一個好的閾值標準看似不難，但當考慮到像是「公平性」這種敏感條件時，問題就變得格外棘手。在圖二中，藍色和橘色族群的信用指數分布有些許不同，究竟要選擇什麼閾值才能避免核准系統歧視任何一個族群？如果站在租貸公司的角度，單純要讓兩個族群的效益達到最大，很顯然地藍色族群的閾值需要大於橘色族群，但藍色族群中就會有較多守信的民眾無法成功貸款，這樣租貸公司該如何向民眾交代機器學習對於不同族群的「公平性」呢？



(圖三左、統一閾值的統計圖表，圖三右、人口均分的統計圖表)

那如果站在客戶的角度討論這個問題，最直覺的方法顯然是統一每個族群的閾值，可是這真的會達到公平的效果嗎？由圖三左可見，整體而言橘色族群獲得較少的借貸機會，因此這個方法沒有辦法達到實質上的公平。

如果目標是讓不同的族群擁有相同的借貸機會，人口均分(demographic parity)會是個值得討論的方法，租貸公司會調整每個族群的閾值，讓「能夠獲得貸款的人」和「族群人口數」的比例維持一致。然而，這可能導致整體看來公平，卻對個體而言不公平。例如，在圖三右中，藍色族群有人保證還款，而且信用指數也比橘色族群中不會還款但入選的人還要高，最後卻因為名額限制而無法取得貸款資格。



(圖四、機會均等的統計圖表)

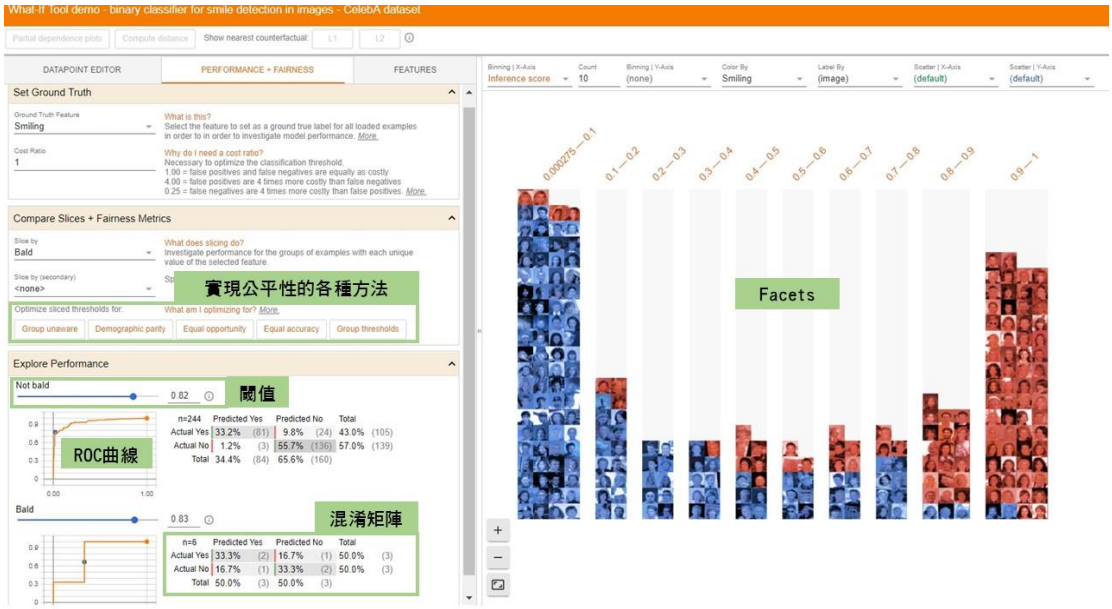
由此可知，我們所追求的「公平」是要保障「值得入選者」的「機會均等(equal opportunity)」。舉上述的例子來說，「保證還款的貸款者」就是所謂「值得入選者」，無論他們身處在哪一個族群，能夠通過評判的機率應該都要相同(如圖四)。儘管這方法可能無法為公司帶來最大的淨利，卻能夠平均每個族群貸款者的品質，進一步消弭差別待遇，擴展公司的經營規模。

What-If Tool

一個好的開發人員在訓練完機器模型後，需要理解模型會不會受資料影響預測結果？對於不同的群體有哪些不同的表現？又會不會因此剝削弱勢族群的權利？而通常開發人員必須要編寫一次性的程式碼，針對特定的目標來測試模型，但這會讓探索機器學習模型的效率降低不少，而且不會寫程式的人也很難參與其中。

為此，Google 在開源的 TensorBoard 網頁應用程式中，推出了 What-If Tool 功能，讓使用者在可以不撰寫任何程式碼的情況下研究模型，只要給 What-If Tool 一個 TensorFlow 模型以及資料庫的指標，該工具就能及時分析，讓使用者一探模型的表現。

What-If Tool 的其中一大特色是可以執行反事實(counterfactual)分析，偵測「如果-那麼」的假設性因果關係。使用者只要按下一個按鈕，就能抓出特徵相似但預測結果不同的資料點進行比較，有效尋找出模型的決策邊界(decision boundary)。另外，使用者也可以手動編輯資料點，以探索資料的變化對模型預測的變化。



(圖五、What-If 的使用介面)

此外，我們也可以利用 What-If Tool 來檢視上述達成公平性的不同方法。除了 ROC 曲線和混淆矩陣會依據使用者選擇的方法變動之外，右方還有 Facets 自動將數據視覺化，使用者也可以點擊上面的資料點以查看特徵，甚至更改數值，讓資料和預測結果之間的關係更清楚明瞭。

參考資料：

1. James Wexler, “The What-If Tool:Code-Free Probing of Machine Learning Models”, Google AI blog, 11 September 2018

2. Moritz Hardt , “Equality of Opportunity in Machine Learning”, Google AI blog, 7 October 2016

3. Martin Wattenberg and Fernanda Viégas and Moritz Hardt, “Attacking discrimination with smarter machine learning”