Delayed Impact of Fair Machine Learning

Author: Asvin Sritharan
Student Number: 1003172684

Professor: Qiang Sun
Due: April 20, 2020

0.1 Problem Statement

Machine learning is applied in many industries which are used in decision making applications while effectively shaping the future of our population. Banks utilize machine learning in deciding who is approved for a loan and who is not. Colleges utilize machine learning in admission and scholarship applications. Machine learning is utilized by advertising agencies in target audience selection. The issue with this application of machine learning is that while it is implemented using data available at the time of creating the model, there is no way to modify the model with changes in population which occur after the implementation of the model. The true harm in this lies in the damage it can cause to marginalized groups.

The difficulty in this problem is that the models result in changes to the population over time. Thus we must adjust the model to its own effects on the population. What follows is the analyzation of circumstances which positively influence the well being of groups which experience greater difficulty and disadvantages compared to non-marginalized groups. This is lead by analyzing the temporal variable of interest which ultimately determines the well being of the marginalized groups.

0.2 Main Result

The significant contribution the paper gives us is the characterization theorem. The characterization is different for DemParity and EqOpt so both of them will be discussed for the sake of clarity in understanding how each classification is achieved. DemParity, known as the demographic parity policy, is a policy used to address differences in two groups by selecting members of a population such that the selection rate of each group is equal. EqOpt, known as the equal opportunity policy, is a policy used to treat people as individuals rather than members of a group such that the conditional probability of selection given the success of a member of a population is not dependent on the success of the population.

Prior to describing the linear programs needed to solve DemParity and EqOpt we must declare $\mathcal{U}(\tau)$:

$$\mathcal{U}(\tau) = \sum_{j\in\{A,B\}} g_j \sum_{x\in X} \pi(x)\tau(x)u(x)$$

$u(x)$ is known as the expected utility function. In the case of a bank, it is the expected return of a loan. If $u(x) = 0$ then this implies that the bank expects to make 0 dollars off of each loan they give out.

To obtain DemParity, we must solve the linear program:

$$\max_{\tau=(\tau_A,\tau_B)\in[0,1]^{2C}} \mathcal{U}(\tau) s.t. \beta = \langle \pi_j, \tau_j \rangle, j \in \{A, B\}$$

To obtain EqOpt we solve the linear program:

$$\max_{\tau\in[0,t_{max}]} \max_{(\tau_A,\tau_B)\in[0,1]^{2C}} \sum_{j\in\{A,B\}} g_j \mathcal{U}_j(\tau_j) s.t. \langle w_j \bigcirc \pi_j, \tau_j \rangle = t, j \in \{A, B\}$$

τ represents selection policies which are chosen by an institution. τ_A represents the policy for group A, similarly τ_B represents the policy for group B. π_A and π_B represents the score distributions for groups A and B respectively. In a bank's scenario, the A score is the credit score of an individual who belongs to group A. A higher score leads to a higher probability of success, a loan approval in the case of the bank.

Looking at DemParity, we realize that we need the same selection rate among both groups. Thus we have β where $\beta_A = \langle \pi_A, \tau_A \rangle$ and $\beta_B = \langle \pi_B, \tau_B \rangle$. We maximize $\mathcal{U}(\tau)$ to obtain the value of $\tau_A == \tau_B$. The resulting τ is the optimal τ for both A and B such that they both enjoy the same selection rate.

EqOpt, as noted before, features a policy of equal opportunity. While the solution for DemParity featured a solution which was parameterized by β, the solution for EqOpt is parameterized by the True Positive Rate: $t = \langle w_j \bigcirc \pi_j, \tau_j \rangle$. w is a weight for the distribution of the corresponding group, that is, if $t_A = \langle w_A \bigcirc \pi_A \rangle$. Depending on the distribution of π_j then w is the value which maximizes the value of t.

0.3 FICO Example

FICO credit scores from the year 2003 were used to create test models and test score change coupled with the application of fairness criteria (Reserve, 2007). These were all done to observe the effects of fairness application on the FICO credit scores. We will analyze the data further.

To begin, it is important to note the FICO data represents the credit scores of two race groups. The fairness criteria seeks to correct the bias of selection against disadvantaged groups by various methods. These methods include DemParity and EqOpt.

0.3.1 The Data

The two race groups in the data are non-Hispanic whites and blacks, which are represented as groups B and A respectively. In this dataset, group A is the historically disadvantaged/marginalized community. The data is obtained from TransUnion using the TransUnion TransRisk scoring system as the method of analyzing scores. The TransRisk system ranks credit from 300 to 850. 300 is the lowest credit score attainable which represents very bad credit, while 850 is the highest attainable credit score meaning excellent credit. The FICO data allows us to estimate the distributions of groups A and B, π_A and π_B. The distribution of the credit scores varies between race and is very apparent [federal research, 80]. This is most noted in the mean scores for each group. The mean TransRisk score of the CDF for non-hispanic whites is 54.0, while the mean TransRisk score for blacks of CDF is 25.6. The original data consists of many different scales for credit scores. Thus it required normalization to make all credit scores to a relatable level. To bring the normalized terms back to that of the original TransRisk scale, we realize that 100 represents a score of 850 and 0 represents a score of 300.

$$850 - 300 = 550$$

$$550 * 0.256 + 300 = 140.8 + 300 = 440.8$$

$$550 * 0.54 + 300 = 297 + 300 = 597$$

This implies that the mean TransRisk score for people of group A is 440.8 while the mean TransRisk score of group B is 597. There is a clear disparity between groups verifying that group A is in-fact the disadvantaged group.

0.3.2 Preparation Before Analyzation

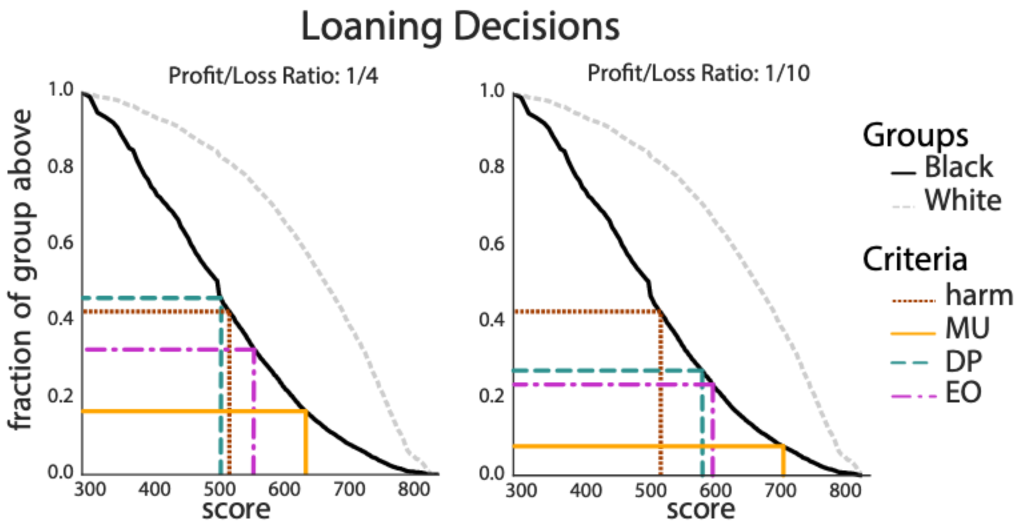
$p_i(x)$ represents success probability for the i'th group. This is calculated by providing penalties to individuals who do not repay a debt while rewarding those who are able to repay their debts. The authors have made the penalty as a drop in credit score by 150 points while the reward is a gain in credit score of 75 points. The importance in this assumption is to regulate individual impact on the credit score of a group.

0.3.3 Analyzation

The analyzation of the data is ultimately done through the cumulative distribution function of the different groups. The authors have chosen two bank utility scenarios to test selection rate impact upon. Namely, ratios $\frac{u_-}{u_+} = -10$ and $\frac{u_-}{u_+} = -4$. Both, methodologies and reasoning were not given as to why or how these values were chosen as utility ratios for the bank.

Digging deeper into the loss/profit ratios, we notice that they are a result of changes to the bank utility. This is not explained either leaving the reader unsure as to how relatable this analyzation and implementation of the model is in real life.

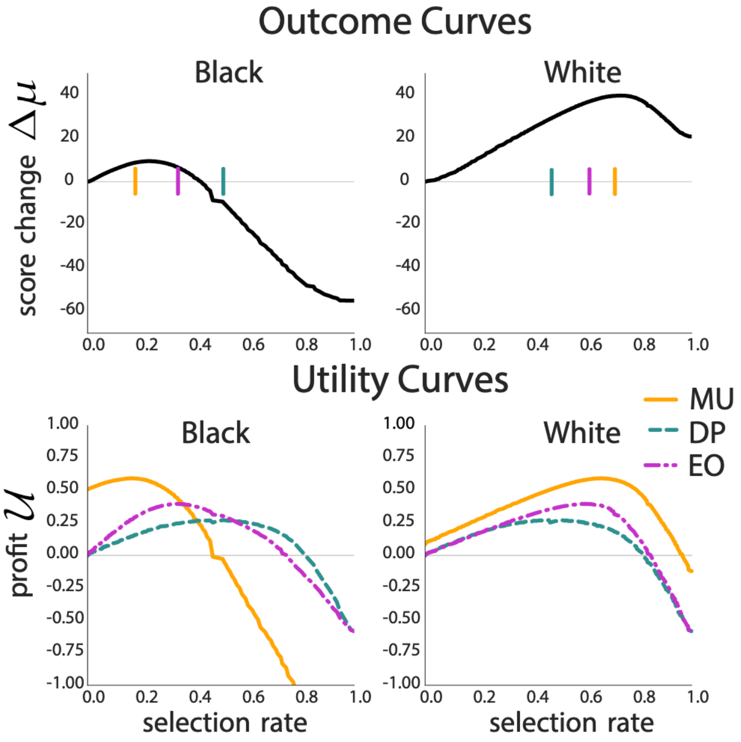
Proceeding to the effects of DemParity, EqOpt and MaxUtil, three policies are graphed respective to the corresponding chosen profit/loss ratio. The MaxUtil policy, known as the maximum utility policy, refers to a policy which is free and is only focused on utility, not demographics or any groups. The main motive is utility, nothing more.



Analyzing the loaning decisions graph (Liu et al. 2018, 21), the profit loss ratio plays an important factor in the effects of loaning decisions in the sense that the impact of DemParity, EqOpt, MaxUtil are dependent on the ratio. The plots are empirical CDFs, plotting score against the fraction of the group above. It is important to note that Group B, the non-hispanic white group, features a more convex curve while the curve of group A is more concave. Thus, the decision thresholds measured, are decided on using the curve of Group A, the disadvantaged group.

The decision boundaries are in similar positions across profit/loss ratios relative to each other. Notably, all positions moved down when switching the ratio from -4, -10. However the decision boundary of DemParity shows a significant difference between ratios. The intersection is formed at a credit score of 500 for a ratio of -4, while the intersection is formed at a credit score of 590 for a ratio of -10.

It is important to note the position of the decision boundaries relative to the harm criteria. Achieving a decision boundary with a score lower than that of the harm criteria indicates the selection rate is overloaning. That is, people with underqualified credit scores are achieving loans which they should not be receiving. Looking at the DemParity boundary, it seems apparent that at a profit/loss ratio of -4, the DemParity decision boundary over-loans to people of Group A, which is counter intuitive and less advantageous to the group due to there being more people who would not be able to repay the loan. EqOpt and MaxUtil both do not cause harm in the -4 or -10 profit/loss ratio meaning one of these options would be the more likely candidate for the selection rate of the bank.



Moving onto the outcome and utility curves (Liu et al. 2018, 22) we are able to analyze which selection rate is the most beneficial for the bank. The outcome curve focuses on the changes of credit scores of a given group depending on varying loaning rates. This variable is represented by β. This β is calculated by solving the linear equations of DemParity, EqOpt and MaxUtil. The value of $\beta_{MaxUtil}$ is the loaning rate which does not differentiate between groups. Rather, it calculates a loaning rate with the sole intention of achieving maximum utilization for the bank. Two graphs are given, one for each group. The characteristic function for DemParity, as mentioned above is:

$$\max_{\tau=(\tau_A,\tau_B)\in[0,1]^{2C}} \mathcal{U}(\tau) s.t. \beta = \langle \pi_j, \tau_j \rangle, j \in \{A, B\}$$

The EqOpt characteristic function is:

$$\max_{\tau \in [0, t_{max}]} \max_{(\tau_A, \tau_B) \in [0,1]^{2C}} \sum_{j \in \{A,B\}} g_j \mathcal{U}_j(\tau_j) s.t. \langle w_j \bigcirc \pi_j, \tau_j \rangle = t, j \in \{A, B\}$$

Noting that the utility of the function is calculated by the function $\mathcal{U}(\tau)$ for both EqOpt and DemParity, we achieve the values needed to plot both the outcome and utility curves.

The utility curves illustrate the total utility given by $\mathcal{U}(\tau)$ endured by the bank for the given group. τ, as mentioned previously, refers to the selection rate of a group. Thus we are able to plot $\mathcal{U}$ based on a specific selection rate to determine the total utility profit for a bank.

Before proceeding, it is important to note that the utility of the bank is set at $\frac{u_-}{u_+} = -4$ and the change ratio, that is the gradual change in credit score is set at $\frac{c_-}{c_+} = -2$. The purpose of this is to demonstrate a situation where the interests of the consumer and the interests of the provider are similar. In this banking example, it refers to the idea that the bank has no intention of creating a very large profit at the expense of desperate people. The farther apart the utility and change ratios are, the more devastating the bank's intentions and people's situation become.

The outcome curve shows a black line for the expected score change for a given selection rate. The yellow, magenta and teal lines represent the global maxima for MaxUtil, EqOpt and DemParity respectively. As we can see, the maxima for MaxUtil occurs when the selection rate is about 0.18 for group A and 0.7 for group B. The maxima for EqOpt occurs at 0.29 selection rate for group A and 0.6 selection rate for group B. The maxima for DemParity occurs when the selection rate is 0.49 for group A and 0.46 for group B.

The utility curve shows lines for MaxUtil, EqOpt and DemParity. That is, the profits for the bank given the selection rates with solutions for DemParity, EqOpt and MaxUtil are plotted. We notice that the curve for DemParity is very similar in both group A and group B. This is the reason why the maxima is very similar for both groups in the outcome plots. The curve for EqOpt seems like it is reflected horizontally when comparing between groups. The curve for MaxUtil shows a similar story, however the reflection is not as apparent. Rather, they show two extremes for the different groups.

The ultimate goal for a bank is to maximize profit, thus the ideal selection which maximizes profit is the selection rate which satisfies the MaxUtil utility for the bank. Referring back to the graphs, we notice that in both groups, EqOpt has a global maxima which is closer to the maximum utility selection rate in comparison to the global maxima of DemParity.

The case of $\frac{u_-}{u_+} = -4$ is analyzed since it was apparent that through the loaning decisions graph, the DemParity would cause harm to those of group A. The key takeaway from this choice was that the case of $\frac{u_-}{u_+} = -10$ occurs when there is unconstrained profit maximization with little regard for those taking part of the bank services. This is usually not a real world application unless the bank's profit motivations are in alignment with the individual credit scores. In the typical scenario, it would not be the case if the utility ratio was -10, therefore the case analyzed is a scenario more applicable to the real world.

0.4 Limitations

The paper highlights the important issue of tackling machine learning oversights over time. The theories discussed by the authors highlight various methods of selection, though we don't get to see the effect of their contributions through a timed output. That is, we don't see the effects of machine learning as time or life changes. The main point of the paper is thus weakened as there is no form of testing done which would allow the reader to observe how machine learning effects those of marginalized communities as time goes on.

An experiment conducted with delayed outputs would be a good start at answering the main question. We do not have an insight as to whether their contributions work in the long run or in a way that does not harm the members of the disadvantaged group.

Another limitation of this experiment is the lack of analyzation towards groups of more than 2 groups (Barocas et al. 2017, 122). In society, there is not only one marginalized group, there are usually many. This is important in discussion as it details the different ways members of a community can be marginalized and how that effects them in a machine learning perspective. This paper makes an assumption that only one marginalized group is dealing with the bank. An analyzation done with more than 2 groups is a more likely scenario in real life and it would be interesting seeing how DemParity and EqOpt function when we are dealing with more than 1 marginalized group.

Additionally, this paper works with classification through linear models. It does not consider models of other types of models. Linear models are not the only models applied in the real world, however this paper only focuses on a linear solution for the issue.

0.5 Future Directions

Future directions can be taken through the oversights of this paper. Most importantly, a study should be conducted with a long term time investment. This will allow us to see the true effects of DemParity and EqOpt in the real world.

Another direction which could be taken is discovering a method of fairness when utilizing multiple groups of marginalized people. This will allow us to tackle groups of people who are marginalized in various ways and have different issues working with the system which is being assisted with machine learning.

0.6 Bibliography

- Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. ”Delayed impact of fair machine learning.” arXiv preprint arXiv:1803.04383 (2018).

- Barocas, Solon, Moritz Hardt, and Arvind Narayanan. ”Fairness in machine learning.” NIPS Tutorial (2017).

- Reserve, US Federal. ”Report to the congress on credit scoring and its effects on the availability and affordability of credit.” Board of Governors of the Federal Reserve System (2007).