Abstraction and Prediction Algorithms:

A Harm-Reduction Framework

Adrian Desilvestro

A dissertation submitted to

Auckland University of Technology

in partial fulfilment of the requirements for the degree of

Master of Business (MBus)

2020

School of Economics | Faculty of Business, Economics and Law

Abstract

ProPublica's allegations, that an algorithmic tool used to predict re-offenders is "biased against blacks", met a wave of criticism from the wider community. Researchers have since shown a trade-off between accuracy and fairness, concluding that the risk tool, COMPAS, was not inherently discriminatory. However, in light of ProPublica's objections, a growing body of literature on assessing fairness in machine learning systems has taken flight. Performance criteria combine quantitative and qualitative elements, so users 'preferences' are hard to specify objectively. This study explores a Pareto frontier framework to illustrate the relative model (in)efficiencies that arise in Risk Prediction Instruments (RPIs). The research follows a logistic framework for estimating recidivism risk, and the design parameters include the choice of fairness constraints and the choice of a bin scoring system (the "bin number"). This dissertation presents three experiments where decision-makers can improve performance in their RPIs: (1) improving efficiency through a relaxed version of the constraint, (2) improving efficiency through 'cost-free' constraint implementation, and (3) improving efficiency through a revised scoring system. Each of these properties results in objective Pareto improvements.

Table of Contents

List of Figures

List of Tables

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of candidate

Acknowledgements

I am deeply grateful to Matthew Ryan for your extensive guidance and support as my adviser. Over the years you have guided, motivated and taught me the importance of putting things in perspective, all with a smile. I would also like to thank Peer Skov, for graciously having shared his knowledge and love of econometrics. Perhaps most importantly, I want to thank my mother for the sacrifices she has made to support me in my studies and for her unconditional love over the years. I would not be the person I am today without you. Finally, thank you to all my professors, friends, and family for making me laugh, cry, and think critically over the years. My experience at AUT has been challenging but I am grateful that I chose to embark on the pursuit of knowledge. I acknowledge all who have pushed me to reflect on each of my experiences, and I hope that as a result, I have become a better student, son, and friend.

1 Introduction

Algorithms have permeated almost every aspect of our daily lives. Decisions on loans, jobs, college admission, and insurance are a few of the many choices where outcomes may be determined on the basis of a risk score. Naturally, prediction algorithms cannot be 100% accurate, but the cost of errors increases considerably in high-stakes settings. For example, prediction algorithms used in the criminal justice setting help judges to assess a suspect's likelihood of re-offending. The algorithm assigns a risk score to a defendant, which is then used to guide decisions on bail or sentencing. If machine learning were to mispredict these scores, it could have significant and detrimental impacts on societal welfare.

It is well known that algorithms perpetuate the data that is fed to them – 'they are what they eat'. So, it is not uncommon to find cases where pre-existing bias is propagated and entrenched. One solution to avoid this process of perpetration and propagation of human bias is to constrain a model by some notion(s) of fairness. A constrained model might generally lead to reducing inequalities. However, as expounded below, the solution is typically ambiguous.

Part of the US criminal Justice System's obligation is to represent all defendants fairly, in good faith, and without discrimination. For example, imagine two students with the same attributes but of a different race. A system that represents both groups equally would mean that both students have an equal probability of being admitted into tertiary education. Similarly, an African-American defendant awaiting his pre-trial release should not receive a lower risk score if he changed her/her race to white – hypothetically, of course.

Unfortunately, achieving "fair outcomes" for two different race groups is not as simple as constraining the model with notion(s) of fairness. An economic agent who decides to constrain the model for equity reasons must face the consequence of giving up accuracy of said model. Intuitively, satisfying *all* the relevant fairness criteria compromises the most accuracy. Similarly, a model free from the requirement to meet any relevant fairness criteria – otherwise known as an unconstrained model – outperforms (in an accuracy metric) all other possible arrangements of the otherwise identical model.

Model designers play a critical role in determining individual outcomes. Outcomes are in part decided by the choice of internal parameters that are used in prediction tools. Design parameters include the number and types of fairness notions, and the choice of a bin-scoring system. The common element amongst these choices is that, in one way or another, they each affect accuracy (some more so than others). Therefore, it is essential that a designer carefully considers the trade-offs in each decision and the potential implications on society.

Attitudes to fairness are comprehensive and multi-dimensional. In an ideal world, a user would like to choose the set of constraints that would maximise aggregate welfare to society. However, the 'optimal' solution – if we could define optimal – is both a qualitative and quantitative one. Since performance criteria combine quantitative and qualitative elements, they are difficult to reduce to a single metric. For this reason, this research paper explores a Pareto frontier framework to illustrate relative model efficiencies.

In this framework we consider a *user*, a *designer* and a *planner* of the risk tool. The designer sets and administers all the design parameters, including the bin number and the constraints to impose. The planner evaluates the quality of the tool. It is the planner's utility function that quantifies the level of 'efficiency' of the model.

The design features of the model affect the planner's welfare, which is summarised by some notion of 'efficiency' and quantified in the form of a planner's utility function. The 'model' is a logistic tool for estimating recidivism risk, converted into a 'bin score', and possibly constrained by some fairness conditions. The relevant performance (efficiency) criteria are 'accuracy' – quantified using Log-Likelihood (LL) – and 'fairness' (expressed as various constraints on score variation across races).

In the next section, we show (in three experiments) how planners can improve efficiency of their prediction tools using the Pareto Framework this paper builds upon. Each of the proposed three experiments show how the "improved" versions of the model lead to performances that are *orders of magnitude* superior to their respective conventional versions.

1.1 Experiment (1)

This first experiment is a robustness test of the accuracy gain from relaxing a constraint. This dissertation explores where accuracy gains can be achieved most quickly per unit relaxation of the fairness constraint.

1. Imagine two different permutations of a model; Model A has twice the accuracy of Model B. However, Model A is constrained such that error rates are within ± 0.01 percent between different racial groups, whereas Model B is constrained to have no disparity in error rates. Most planners would not hesitate to choose Model A over Model B – and their choice would be well justified. A gain of 100% accuracy to the model at the expense of allowing some trivial variation in error rates seems sensible and legitimate. If the planner would agree with this proclamation, then the planner's utility function would satisfy $U_A > U_B$. This means that the indifference curve containing Model A is above Model B's.

Distinct views on fairness are qualitative, and thus deviations from a given fairness constraint are hard to quantify objectively. Experiment 1 has a different "flavour" from the two following experiments, which exhibit straightforward Pareto improvements. Experiment 1 suggests that the planner may be willing to give up some 'small' amount of 'fairness' in order to gain significant accuracy of the risk tool. The experiment shows that accuracy may be quite "cheap". If the planner's utility function has a Marginal Rate of Substitution (MRS) for accuracy vs. fairness that is not too large (in magnitude), s/he may find the trade-off worth making. This paper quantifies the local trade-off so that the planner can compare it to his/her MRS. Moreover, distributions of probability estimates illustrate how "tight" a strict version of the constraint can really be, and so provides strong motivation for relaxing them.

1.2 Experiment (2)

Experiment (2) proposes that a designer can improve efficiency in their RPI through 'cost-free' constraint implementation. The following example illustrates this possibility.

2. Imagine an algorithm constrained by a single fairness condition, which we denote as Model A. Now suppose a second condition could be imposed ‘free of charge’ – that is without costing the model any accuracy – we denote this permutation as Model B. Clearly, the second arrangement is strictly better than the former, thus any rational designer would choose Model B.

In this second example, we can see that there is no ambiguity between the choices. One choice is strictly better than the other, that is, $U_B > U_A$. In other words, all else equal, a designer can strictly improve fairness outcomes without costing the model any accuracy; thus, attaining prediction tools of higher efficiency.

1.3 Experiment (3)

A third Experiment describes the bin-number frontier. The resulting Pareto improvement stems from an improvement of model performance in two dimensions: accuracy (higher Log-Likelihood (LL)) and the “informativeness” that a unique bin-scoring system provides. “Binning” refers to the partitioning of continuous variables into categorical groups. This system of transforming continuous variables to discrete variables is often used by designers to discover patterns (which would be difficult to analyse otherwise) and so that the data fits naturally into the framework.

3. Imagine a user who chooses a model with a 2-bin scoring system and constrains the model by a single fairness condition. Now suppose a 3-bin scoring system meets the same fairness condition but achieves a higher LL. The latter model would be an objective improvement in efficiency (improves accuracy with the same type of fairness constraint) and also better ‘informs’ the user about the relative risk of a defendant; the model would give judges a more comprehensive ‘set of scores’ to work with. This dissertation verifies the possibility for both strict and relaxed versions of the fairness constraint.

2 Research questions

The experiments conducted in this paper used ProPublica's dataset to illustrate Pareto improvements (Experiments 2 and 3) or potential welfare-improving trade-offs (Experiment 1) by assessing the relative trade-offs in three dimensions: accuracy (measured in LL), fairness and 'informativeness'. Designing efficient RPIs is highly technical due to the confounding interactions between these three dimensions and their local trade-offs. The experiments this paper sets out illustrates three key findings: 1. It explores where accuracy gains are achieved most quickly per unit relaxation of the fairness constraint. 2. When 'balance for the positive class' (a fairness notion) is satisfied, a designer can impose 'balance for the negative class' without costing the model any accuracy (and vice versa). 3. A 3-bin scoring system delivers a higher LL than a 2-bin scoring system while simultaneously improving 'informativeness'.

A contribution of this work is to provide a theoretical and statistical framework for reasoning about equity in prediction tools. The harm-reduction framework I build in this dissertation aims to minimise the adverse societal effects that stem from unequal treatment based on group membership. Moreover, this dissertation investigates the degree to which prediction is compromised when administering interchangeable design parameters. This article does not 'solve' any treatment asymmetries in its full breadth, but instead highlights the cases where users can improve efficiencies in their existing RPIs and as a result, improve societal welfare.

3 Literature review

In Prediction Policy Problems, the authors illustrate where predictive techniques could lead to; higher utility in patients undergoing joint replacement procedures, in addition to considerable cost-savings for insurance companies (Kleinberg et al. 2015). For example, replacement of affected joints is frequent among the elderly suffering from osteoarthritis (pain and stiffness in bones). Patients who undergo surgery have seen meaningful improvements in quality of life. However, the patients have to first endure a 12-month recovery period after the time of the operation. During these months, it is not uncommon to see significant pain and disability (disutility) inflicted on the patients. Therefore, funds should not be allocated toward the *riskiest* patients since these surgeries would have a higher probability of being futile. If capital was allocated toward patients with a higher probability of surviving the recovery period (and the previous surgery), then accrued benefits could be on a potentially large scale.

Kleinberg et al. drew a 20 percent sample of 7.4 million Medicare beneficiaries in the United States, and from these, the authors found that 4.2 percent of eligible patients would die within 12 months of surgery, and 1.4 percent die from complications during the operation. This seems to suggest that, on average, a low fraction of the patients who undergo surgery end up with a futile outcome. However, the result is misleading since the policy decision is really about "whether surgeries on the predictably riskiest patients were futile" (2015, p.493). Since the payoff (utility) for undergoing surgery depends on the event of mortality, the application becomes a pure prediction problem and *not* a problem for causal inference. The authors simulated the benefits using a regularised logistic regression trained on 65,395 Medicare beneficiaries undergoing joint replacement. They concluded that 10,512 deaths could have been averted and that cost-savings could be upwards of USD 158 million per year if predictive methods were employed to identify the risk level of a patient.

Predictive methods, such as a logistic regression, could be employed in many domains which could lead to larger societal benefits. The challenge is identifying the domains in which prediction tools can be applied and expanding our conceptual understanding of ‘what is predictable’.

3.1 ProPublica analysis

In 2017, the New York Council passed the first Bill (Algorithmic Accountability Bill) in the United States to address algorithmic discrimination in government agencies (Kirchner, 2017). The Bill was designed to make the city’s algorithms fairer and more transparent. Following the Bill, an article published by a non-profit New York journalism company ProPublica revealed that an RPI called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was inherently biased against African-Americans (Angwin et al., 2016). The algorithm was developed to increase the efficiency of bail and sentencing decisions for the US criminal justice system and was a widely used risk tool by courts to predict a defendant’s probability of recidivism.

A general algorithmic pipeline maps characteristics of an individual to a system architecture which generates a probability estimate of some event occurring (e.g., a defendant recidivating) (Fig. 1). These probability estimates are assigned to a corresponding ‘bin’ that then becomes ‘useful’ and measurable to the user (e.g., a generated “risk score” on a defendant’s relative level of risk). The COMPAS algorithm used a range of scoring integers between 1 and 10: a score of 10 representing a defendant with the highest likelihood of re-offending and a score of 1 representing a defendant with the lowest likelihood of re-offending. The company classified a defendant as “high risk” if the algorithm assigned the individual a score above four and classified a defendant as “low risk” if the generated score was equal to or below four (i.e., a high risk “threshold” value of 4).

The company’s stylised findings were that prediction “fails differently” for black and white defendants. They found that African-American defendants who did not subsequently re-offend had higher average risk-scores than white defendants who also did not subsequently re-offend. Similarly, white defendants who subsequently re-offended had lower average risk-scores than their African-American counterparts.

Their primary finding was that African-American non-recidivists were almost twice as likely to be misclassified as high risk compared to white non-recidivists, while white recidivists were significantly more likely to be misclassified as low-risk compared to African-Americans recidivists.

Alex Chouldechova (2017), a leading researcher in the field, studied "imbalance in misclassification rates between group membership" in the COMPAS algorithm. She defined the asymmetrical treatment between the two groups as an imbalance in "false-positive rates" (FPRs) and "false-negative rates" (FNRs) between blacks and whites. An FPR refers to the ratio between the number of negative instances (for a subgroup of individuals) wrongly categorised as positive and the total number of actual negative instances (for the same subgroup). Similarly, the FNR is calculated as the ratio between the number of positive instances (for a subgroup of individuals) wrongly categorised as negative and the total number of actual positive instances (for the same subgroup). In the context of predictive policing, a positive instance is the event where a defendant recidivates within the next two years and a negative instance is the event where a defendant did not subsequently recidivate.

Chouldechova (2017) found that COMPAS has "considerably higher FPRs and lower FNRs for black defendants than for white defendants" (2017, p.2). Further, she proved that this was not only true at a high risk cut-off threshold was equal to four (the threshold ProPublica used in their analysis), but also across *all* cut-off values. Chouldechova verified that the COMPAS tool had satisfied the fairness notion "predictive parity" (when the likelihood of recidivism among high risk offenders is the same, irrespective of group membership). Furthermore, she illustrates that predictive parity is mutually incompatible with error rate balance[1] when recidivism prevalence[2] differs across group membership.

[1] *Error rate balance* is when the both the notions of FPRs and FNRs are simultaneously equalised across race.

[2] *Prevalence* is the proportion of individuals who recidivate in a given population.

3.2 Northpointe rebuttal

Northpointe, the developers of the COMPAS tool, firmly rejected the ProPublica's allegations; that COMPAS was biased against African-Americans. Northpointe asserts that COMPAS was (and still is) equally fair to black and white defendants (in the sense of being *well-calibrated*). The company stipulated that the risk tool was devised to achieve the notion of predictive parity and they had executed on that objective. "A test that is correct in equal proportions for all groups cannot be biased" (Angwin and Larson, 2016).

Despite the criticism that ProPublica had met from Northpointe, their fundamental claim is not inherently erroneous. There should not be such a disproportionate number of black defendants misclassified as 'high risk' compared to their white counterparts. Similarly, there should not be such a disproportionate number of white defendants misclassified as 'low risk' compared to their black counterparts. However, the fact that recidivism prevalence had differed between African-Americans and Caucasians in ProPublica's case is where this challenging goal emerges: designing an architecture that is both equally accurate and equally fair (or at least closer to that goal) to blacks and whites.

3.3 Kleinberg, Mullainathan and Raghavan

The literature shows that there is no simple, all-encompassing notion of fairness for this class of systems. The question of how a risk tool can be both fair (satisfy predictive parity) and unfair (violate error rate balance) at the same time attracted top researchers across the world. Kleinberg – a computer science professor at Cornell University –and his co-authors, found that satisfying balance for the positive and negative classes (or "group fairness") while simultaneously satisfying calibration within groups, is mathematically impossible to achieve when recidivism prevalence differs across group membership (Kleinberg, Mullainathan, and Raghavan, 2017). These definitions of "group fairness" and "calibration within groups" are formalised in Chapter 5.2.

3.4 Race-aware algorithms

Since the release of the ProPublica article, policymakers had begun to expose their risk-tools to "race-blindness" in the belief that it would eliminate human bias in sentencing decisions (Kleinberg et al. 2018). Race blindness occurs when a user excludes race variables (known as *protected* or *sensitive* features) from their risk tools. The authors show that race-blindness may in fact be doing "more harm than good." They use nationally representative US data to predict the likelihood of college success in the use of student admission decisions. Furthermore, they use a model to show than an equitable planner may be able to increase equity and "efficiency" by including protected variables rather than omitting them. It is important to distinguish the difference between how Kleinberg et al. defines "efficiency" and how this dissertation defines the term. This dissertation uses the term "efficiency" to describe the overall performance of the model (on the accuracy, fairness and informativeness dimensions together), as captured by the planner's utility function. Kleinberg et al. uses the term to illustrate the model's relative performance in only one dimension, accuracy.

Kleinberg et al. (2018) define two types of welfare planners: the efficient planner and the equitable planner. The efficient planner cares only about the predicted performance of the students whereas the equitable planner cares about both student performance and racial composition. For example, an equitable planner might choose to increase the fraction of black students admitted by lowering the high-school SAT score threshold for only minorities. The authors show that when a planner includes protected variables in his/her model, both the equitable and efficient planner are strictly improving their respective objective functions (provided that the protected variable is actually useful in predicting students' college success).

The authors used a public time series dataset called "Department of Education's National Education Longitudinal Study of 1988" (NELS:88) to capture students' GPAs who entered eighth grade in 1988 and tracked (GPA) progress into their mid-20s. They discovered that the "race-aware" algorithm outperformed the race-blind model. For any given level of diversity, the fraction of students selected by the race-aware algorithm who achieved high grades (GPA > 2.75) exceeded that of the race-blind algorithm. In contrast, the race-blind algorithm led to the highest fraction of selected students receiving low grades (GPA < 2.75).

Since the efficient planner is only concerned about the ranking of the outcome, having a race variable included is always the best choice. Even the equitable planner (one who might lower the threshold for minorities) should include a race variable.

4 Data

Data was retrieved from an online public data repository, “Github Inc.” (uploaded by ProPublica). The data sample includes observations of 7,215 criminal defendants from Broward County, Florida, across the years 2013 and 2014. The file contains data for each defendant on age, sex, race, number of prior offences, charge degree, in addition to a few generated variables: violent risk scores, and COMPAS risk scores. Table 1a reports the summary statistics of age, number of prior offences, decile scores and violent decile scores. Violent risk scores are risk assessments on whether a defendant will commit a violent crime. Based on the severity of the criminal cases, the Supreme Court of Florida classifies charge degree into two broader categories; misdemeanours and felonies. Misdemeanours charges are generally considered less severe and are further classified into first- and second-degree crimes. Felonies are considered severe and are divided into five categories including a separate category for drug-related offenses. The sample also contained records on two-year recidivism outcomes for each defendant which was the metric used to derive our findings.

Since “charge degree” was an ordinal variable, this dissertation transformed charge degree into two groups of binary variables. The groups were categorised by “felonies” and “misdemeanours and other”. Table 1b reports that there were a total of 1,006 Felonies, of which 67.2% were committed by African-American defendants and 32.8% by Caucasian defendants.

The data was narrowed down to 6,150 unique appearances and of these: 3,696 were African-American (r=b) and 2,454 were white (r=w); 2,867 were recidivists (Y=1) and 3,283 non-recidivists (Y=0); 4931 male defendants (s=m) and 1,219 female defendants (s=f) (Table 2).

Table 2 reports proportions and average scores from three variables, recidivism outcome, race and sex, or any combination of the three (one outcome can be conditioned against either one or both of the corresponding variables). For example, imagine a defendant was drawn at random from the “recidivating” sub-population. Table 2 would report there would be an 84.4% chance of a male recidivist being drawn and a 66.3% chance of an African-American being drawn. Applying the same process, we can, for example, also condition ‘race’ on two variables. For example, if a defendant was drawn at random from the “recidivating *and* male” subgroup, there would be a 68.3% chance of an African-American male recidivist being drawn.

Another noteworthy observation is that African-Americans have risk scores that are on average 1.44 times higher than Caucasians (Table 2). Furthermore, Fig 2 and Fig 3 illustrate how the distribution of these risk scores are skewed toward low-risk brackets for the white group while the distribution of risk scores for African-Americans is uniform across all levels of risk.

5 Prediction algorithm

This research paper engineered an original and novel architecture using Microsoft Excel and "@Risk Solver" software for constrained optimisation. The system runs iterative solutions that predict recidivism.

To begin the design process, defendants are classified as positive or negative instances of a given property. A positive instance is where (s)he possesses the relevant property while a negative instance is where (s)he does not possess the property in question. In the criminal justice setting, a positive instance is where the defendant recidivated (Y=1). Similarly, if the defendant never recidivated, then the (s)he is referred to as having a negative instance (Y=0).

To estimate the probability that a defendant will recidivate, we start with the following expression:

$$P = Pr(Y = 1|X_i) \qquad 1$$

That is the probability that an individual with attributes X_i will be a positive instance. We denote X by the vector for all X_i such that $X = [X_1, X_{2,\cdots,}X_J]$ and $J = n_w + n_b$ where n_w denotes the total number of white defendants n_b, the total number of black defendants. A logistic framework was chosen as it yields well to such measures of deriving probability estimates and corresponding risk scores. Using the logistic formula, the log-odds is a linear function of X_i:

$$ln\left(\frac{P}{1-P}\right) = \sum_k \beta_k X_{ki} \qquad 2$$

$$\frac{P}{1-P} = e^{\beta_0+\beta_1 X_{1i}+\cdots+\,\beta_k X_{ki}} \qquad 3$$

$$P(\beta, X_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\cdots+\,\beta_k X_{ki})}} \qquad 4$$

That is the probability a defendant recidivates conditional on X_i. Note that the estimate is not binary but instead a continuous probability estimate where $P \in (0,1)$.

The primary goal in mind when devising a risk tool is to predict future offenders and non-offenders as accurately as possible. Having this objective in mind, we solve for β values that maximise the following function:

$$\prod_{i \epsilon R} \left[\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}}\right] * \prod_{i \epsilon NR} \left[1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}}\right] \qquad 5$$

R denotes the subset of observed recidivists, and NR denotes the subset of non-recidivists. I refer to the product of the $i \in R$ terms as the LHS of expression (5) and the product of the $i \in NR$ terms as the RHS of the expression. The LHS is the product of all recidivism probabilities for the subset of defendants who went on to recidivate. The RHS is the product of all non-recidivism probabilities for the subset of all defendants who did not recidivate.

Expression (5) is the probability of observing everyone in R re-offending and everyone in NR not re-offending. As its value approaches 1, the function begins to reflect (with increasing accuracy) the likelihood of the recidivism pattern described by the sets R and NR. Therefore, our objective is to find the solution vector $\hat{\beta} = \left[\hat{\beta}_0, \hat{\beta}_{1,\cdots,}\hat{\beta}_J\right]$ that maximises the above function. It is conventional to work with the log transform of equation (5), so we maximise the following function, denoted $F(\beta, X)$:

$$\sum_{i \in R} ln\left[\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}}\right] + \sum_{i \in NR} ln\left[\frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})}}\right] \qquad 6$$

5.1 Bin and score assignment process

The methodology in this paper of mapping probability estimates to risk bins follows a widely used convention in this literature. The convention arises because scores are simpler for users of the risk tool to work with. The probability estimate $P(\hat{\beta}, \sigma)$ for an individual with attribute vector $X_i = \sigma$ is mapped to a *risk bin*, $\theta_k\left(P(\hat{\beta}, \sigma)\right) \in \{1, 2, \ldots, k\}$, where k denotes the number of bins. Each bin has equal 'probability width'. That is, $\theta_k(p) = 1$ if $0 \leq p \leq 1/k$ and $\theta_k(p) = i > 1$ if $(i-1)/k < p \leq i/k$. For example, in a 10-bin scoring system, $\theta_{10}(p) = 1$ if $0 \leq \rho \leq 0.1$, $\theta_{10}(p) = 2$ if $0.1 < \rho \leq 0.2$ and so forth.

Risk scores are assigned for each defendant, based on the bin number they are partitioned into. The following equation denotes a defendant's risk score, where S_σ^k is a normalised value of $\theta_k\left[P(\widehat{\beta}, \sigma)\right]$:

$$S_\sigma^k = \frac{\theta_k\left[P(\widehat{\beta}, \sigma)\right]}{k}$$

For example, in a 10-bin scoring system, if $P = 0.45$, this gets mapped to bin 5 and hence to a 'normalised' score of 0.5.

5.2 Formalising fairness

Accurately predicting future recidivism is the embodiment of any probabilistic classification in sentencing decisions and is a central objective for any planner. However, in a world with an ever-widening gap in inequality, a shifting emphasis on the creation of fair outcomes for different groups is taking precedence.

To facilitate comparison with previous work, this paper revisits three pre-existing fairness notions; *balance for the positive class*, *balance for the negative class* and *calibration within groups*. Fairness notions are designed with the purpose of reducing prejudice from risk tools. Determinations must be non-discriminatory and fair in sensitive features, such as race and gender. Each notion addresses the relative treatment of different demographic groups so that one group is not systematically favoured more than the other. The group of defendants with positive instances is the *positive class* and the group of defendants with negative instances are the *negative class*.

5.2.1 Class balance

The first fairness constraint we explore can be expressed in the equation (A), below. The average (expected) risk score assigned to the members of the African-American group ($r = b$) for the positive class ($Y = 1$) should equal the average (expected) risk score assigned to the members of the Caucasian group ($r = w$) who also belong to the positive class, where r denotes race. This is known as balance for the positive class and may be formalised as follows:

$$E[S_\sigma^k \middle| Y = 1, r = b] = E[S_\sigma^k \middle| Y = 1, r = w] \qquad A$$

The second fairness criteria, expressed by the equation below (B), is defined as balance for the negative class. The average risk score assigned to the members of the African-American group ($r = b$) for the negative class ($Y = 0$) should equal the average risk score assigned to the members of the Caucasian group who also belong to the negative class.

$$E[S_\sigma^k \middle| Y = 0, r = b] = E[S_\sigma^k \middle| Y = 0, r = w] \qquad B$$

The COMPAS risk tool incessantly assigned African-Americans from the negative class higher-risk scores than Caucasians who also belonged to the negative class. Similarly, the risk tool assigned Caucasians from the positive class lower risk scores than African-Americans who also belonged to the positive class. Thus, the imposition of fairness constraints in (A) and (B) aim to ameliorate this discrepancy of false findings between the two groups. However, as detailed in the prior sections, the enforcement of these fairness constraints are not inherently 'free' due to complex trade-offs.

5.2.2 Calibration within groups

A third (fairness) constraint which a planner may choose to satisfy is that of a model that is free from predictive bias. *Calibration within groups* fits this definition and has been a widely accepted and adopted empirical fairness assessment. Calibration within groups requires that a score $S_\sigma = s$ reflects the same probability of recidivism irrespective of the individuals' race group membership. That is, the following is true for all values of $s \in \left\{\frac{1}{k}, \frac{2}{k}, \dots, \frac{k}{k}\right\}$.

$$\rho[Y = 1 \big| r = b, S_{\sigma}^{k} = s] = \rho[Y = 1 \big| r = w, S_{\sigma}^{k} = s] \qquad C$$

The rest of this paper focuses on (A) and (B) since only two of the three fairness constraints can be imposed simultaneously (given Kleinberg et al.'s result). This decision to focus on (A) and (B) is made for convenience. A limitation for excluding constraint (C) is that we do not get to directly observe the 'violation' to (C) when satisfying conditions (A) and (B) simultaneously. It may be possible that experiments conducted in this paper might have very different outcomes (in efficiency rankings) if we were to include a notion (C). Studying the interaction(s) between a (C) constraint with all other design parameters (e.g., fairness constraints, LL or a ROC metric for accuracy and a bin-scoring system), would make for an exciting future research challenge.

5.3 Verifying Chouldechova's results

Findings indicate that COMPAS is indeed well-calibrated under a 10-bin scoring system (Fig 4). Furthermore, the COMPAS tool satisfied the fairness notion of predictive parity (Fig 5). ProPublica's assertion that "FPRs are approximately two times higher for blacks and FNRs (error rates) are almost two times higher for whites", by this paper's analysis, is accurate when the cut-off (denoted by (S_{HR}) is equal to four as in ProPublica's analysis: this is illustrated by Fig 6 and Fig 7. Moreover, imbalances in error rates are significantly different between group membership at *any* cut-off threshold: highlighting the incompatibility between error rate balance and Predictive parity.

6 Experimental Results

The previous sections outlined that predictive classifiers cannot satisfy all ‘dimensions’ of accuracy, fairness and informativeness simultaneously. The central design issue is how to initialise payoff functions for each of the performance criteria (dimensions) so to maximise societal welfare. We conducted preliminary investigations of assessing relative model performances, as captured by the planner’s utility function. Since we cannot directly observe the planner’s utility, this dissertation (i) focuses on identifying Pareto improvements in experiments 2 and 3 and (ii) quantify the accuracy-fairness trade-off from a particular relaxation of fairness constraints in experiment 1.

6.1 The unconstrained model

The unconstrained model is a reference or ‘benchmark’ to quantify trade-offs when imposing different constraints. The unconstrained model predicts recidivism using all 12 non-race variables (Table 3) and imposes no fairness constraint. Intuitively, the unconstrained model achieves the highest predictive accuracy out of all possible model variants with a resulting LL of $F(\hat{\beta}, X) = -726.15$. Fig 18 illustrates the distribution of probability estimates from the unconstrained model. We observe that the probabilities for the recidivists (Y=1) and non-recidivists (Y=0) are accurately distributed by the model for both black and white defendants.

6.2 The first experiment

This first experiment evaluates the logistic tool’s accuracy gain by imposing a relaxed version of the constraint(s). We first examine the extent to which each constraint (A) and (B) reduces LL relative to the unconstrained model (our preliminary benchmark) and then measure the accuracy gain from relaxing these “strict” constraints. Furthermore, we compare probability distributions between the relaxed and strict versions of the constraint to motivate this discussion on why a planner should choose a relaxed version (over the strict version) of the constraint.

Each constraint (A) and (B) is expressed as an equality, which is the ‘strict’ form. A strict constraint is non-negotiable and binds the algorithm to meet the condition at all costs. In this context, constraints (A) or (B) are satisfied when differences in expected risk scores between black and white defendants (LHS-RHS difference) is precisely zero. Similarly, for each constraint, we could think of a ‘relaxed’ form which imposes an upper and lower bound on the LHS-RHS difference, perhaps after some ‘normalisation’ of this difference. A relaxed constraint means that the logistic formula can operate within the user-specified upper and lower bounds to compute maximum log-likelihood.

The relaxed version imposed on the risk tool in this paper follows an unpaired two-sample t-Test to define the upper and lower bounds. An unpaired t-Test tests the null hypothesis such that the population means related to two independent, random samples from an approximately normal distribution are equal (Armitage and Berry, 1997). A t-stat less than 1.96 suggests that differences in conditional distributions are not statistically significant at the 5 percent level. The two-sample t-Test formula is as follows:

$$t = \frac{\bar{S}_w^k - \bar{S}_b^k}{\sqrt{\left(\frac{\sigma_w^2}{n_w} + \frac{\sigma_b^2}{n_b}\right)}}$$

The equation above tests the differences in the underlying distributions of average risk scores between black and white defendants. The test was carried out on the positive class if a relaxed version of (A) was imposed, and on the negative class if a relaxed version of (B) was imposed. Let w denote the white defendants and b denote the African-American defendants. The sample mean risk scores for any chosen ‘k bin number’ is denoted by $\bar{S}_w^k$ and $\bar{S}_b^k$. The sample variances are σ_w^2, and σ_b^2, the sample sizes are n_w, n_b.

6.2.1 The ‘strict’ results

We begin with some notation for convenience purposes. Let $\hat{F}_k^{\omega}(\Omega)$ denote the corresponding constrained maximum LL for the model constrained by a k −bin, scoring system, where $\omega \in \{\alpha, \gamma\}$ and $\Omega \in \{A, B, Z\}$. The constraints can come in the form of a strict version, α, or a relaxed version, γ. Ω denotes whether constraint (A) or (B) or (Z) is imposed across the model and Z corresponds to satisfying the joint conditions of (A) and (B) simultaneously. Moreover, we abbreviate a model constrained by the foregoing design parameters in this form: “$\omega\Omega k$”. For example, $\alpha Z9$ corresponds to a model constrained by a strict version of (Z) and using $k = 9$ bins as a scoring system.

The following example illustrates the maximum log-likelihood for all models constrained by $\alpha\Omega 10$. A bin-scoring system of $k = 10$ is used to make the comparison since it replicates the scoring system from the ProPublica dataset. Table 4 reports that model constrained by $\alpha A10$ computes a maximum LL, $\hat{F}_{10}^{\alpha}(A) = -3793.09$. Similarly Tables 4 and 6 reports that the models constrained by $\alpha B10$ and $\alpha Z10$ compute a maximum LL, $\hat{F}_{10}^{\alpha}(B) = -3834.17$ and $\hat{F}_{10}^{\alpha}(Z) = -4076.82$, respectively. Thus $\hat{F}_{10}^{\alpha}(A) > \hat{F}_{10}^{\alpha}(B) > \hat{F}_{10}^{\alpha}(Z)$.

The underlying probability distributions for the three models constrained by $\alpha\Omega 10$ are reported in Figs 14 - 17. To better understand how the model is assigning the probability estimates to defendants, we partition the probabilities by outcome (Y) and race. We observe that all probabilities for the model constrained by $\alpha Z10$ are confined within the range of 0.4 and 0.5 (Table 12 reports summary statistics for these probabilities). The story unfolds from here. The pattern we observe in Figs 14 – 17 (strict constraints) is that as k decreases, the probability distributions between the two predicted classes (recidivists and non-recidivists), diverge. Thus, a decreasing k implies that our logistic tool can predict recidivists and non-recidivists more accurately. We discuss the relationship between k and LL in more detail in experiment 3.

6.2.2 The 'relaxed' results

This section highlights why planners should consider using the relaxed version of a constraint over the strict version. Consider the following arbitrary example. Tables 7 – 10 report that the three models constrained by $\gamma\Omega10$ compute the same maximum log-likelihood such that $\hat{F}_{10}^{\gamma}(Z) = \hat{F}_{10}^{\gamma}(A) = \hat{F}_{10}^{\gamma}(B) = -1161.90$. Table 10 reports that the LHS-RHS difference in $\hat{F}_{10}^{\gamma}(A)$ produces |T| = 1.526 and the LHS-RHS difference in $\hat{F}_{10}^{\gamma}(B)$ produces |T| = 1.959. Moreover, Table 10 reports that a model constrained by $\gamma\Omega k$ for any constraint Ω and k value sampled in this paper, satisfies the condition |T| < 1.96. Thus risk scores are not statistically significantly different between the LHS and RHS at the 5 percent level (constraint (Z) is, of course, satisfied if both (A) and (B) are). This suggests for a partial relaxation of the constraint, accuracy of the prediction algorithm improves by over 300 percent.

We can glean a lot of information by examining the probability distributions for each of the constraints. Figs 10 - 13 exhibit the distributional probability estimates for all relaxed versions of the model. By examining these distributions, we observe how the risk tool "searches" for probabilities close to zero when Y=0 and close to one when Y=1 subject to the Ω constraint. The patterns we encounter from the probability distributions in the relaxed versions of the constraint(s) resemble the same pattern generated by the corresponding estimates from the unconstrained model. In fact, this dissertation makes the striking discovery that a model constrained by $\gamma\Omega9$, "finds" probability estimates that are not *statistically significantly different from the estimates generated by the unconstrained model*. To be sure, this paper t-Tests the probability estimates generated from the model constrained by $\gamma\Omega9$ against the corresponding estimates generated from the unconstrained model. The result: a t-stat of 1.72, suggesting that the probabilities are not statistically significantly different at the 5 percent level (Table 13).

6.2.3 Comparative assessment

We compare the qualitative difference (expressed in LL) and quantitative difference (expressed in fairness and informativeness) between the relaxed version of the constraint and: (1)

unconstrained benchmark; (2) the strict version. Let $\hat{u}_k^{\omega}(\Omega)$ represent the optimal payoff utility for a planner, where are ω, Ω and k represent the same parameters that were defined the maximum LL function $\hat{F}_k^{\omega}(\Omega)$. The utility function represents the maximised 'efficiency' of the model (performance in the accuracy, fairness and informativeness dimensions together).

This first experiment highlights the accuracy gain from relaxing the constraint. A user can think of $\hat{u}_k^{\omega}(\Omega)$ as a function of the constraint slackness level |T|. A planner with a well-defined MRS between |T| and accuracy (LL) could be compared to a trade-off rate in order to find the optimal compromise. Since the MRS will change according to differences in planners utility functions, we cannot say whether or not a particular planner will find the trade-off worth making.

The accuracy gain from relaxing a constraint, however, is quite dramatic. Fig 10 and Fig 14 illustrate the differences in how these probability distributions (between strict and relaxed versions of the constraint) are formed and help to visualise really how tight these constraints are. The implied distribution of risk scores is much more plausible under the relaxed constraint and gives strong motivation for relaxing them.

6.3 The second experiment

This section expounds on how a designer can improve efficiency of their RPIs through 'cost-free' constraint implementation. We saw in the last section that $\hat{F}_{10}^{\gamma}(Z) = \hat{F}_{10}^{\gamma}(A) = \hat{F}_{10}^{\gamma}(B)$, thus our second experiment illustrates that a designer who chooses to constrain a model with the relaxed constraints of (A) or (B) can systematically satisfy a relaxed version of (Z) without costing the model any accuracy and/or any informativeness. In other words, any model constrained by γZk computes the same maximum LL as a model constrained by γAk or γBk and this is true for any k value that we sampled in this paper $k \in \{1, 2, 9, 10\}$ (Tables 7, 8 and 9). Indeed, this Pareto improvement does not transcend to a *strict* version as reported in Tables 4 – 6. This strict improvement in fairness, at no cost to accuracy or informativeness, implies that a Pareto improvement is possible if a model is operating under the constraints γAk or γBk. The resulting gain in efficiency (planner utility) can be expressed as follows: $u_k^{\gamma}(A) = u_k^{\gamma}(B) < \hat{u}_k^{\gamma}(Z)$.

6.4 The third experiment

This final experiment illustrates why choosing a higher bin-scoring number (k-value) is more informative to a user and why we might expect a trade-off between accuracy and bin number, for a given fairness constraint. Findings illustrate the unique possibility of a Pareto improvement through informativeness *and accuracy* dimensions for a given fairness rule.

To show why a higher k-value is more informative, imagine a model constrained by a one-bin scoring system. All probability estimates are assigned to the same risk bin, θ_1, thus the average normalised score also is equal to 1. Since there is no variation in risk scores, what would a judge do with such a list – where all defendants were assigned the same risk score, $S_\sigma = 1$? This example highlights the problem with choosing a low-numbered bin-scoring system: Fewer bins are less informative to the user about a defendant's level of risk.

As k approaches infinity, the bin-constraint requires average estimated recidivism probabilities to coincide across groups. For example, imposing (A) with $k = \infty$ would be the same (i.e., produce the same parameter estimates and the same maximised LL) as imposing a constraint which required the average estimated recidivism probability across white recidivists to equal the average estimated recidivism probability across black recidivists.

6.4.1 The pattern of a higher k

From the first experiment, we saw how "confining" a strict constraint can be – due to how the logistic tool distributes its probability estimates across classes (positive and negative). As we move from $k = 10$ to $k = 2$ (Figs 14 – 17), we observe how the probability distributions of Y=1 and Y=0 diverge. This 'divergence' corresponds to an improving maximum LL as the model can predict the "correct" outcomes for the two classes more accurately.

As we discover, there is one *consistent* "violation" to this trend (of a decreasing k for an increasing LL) when imposing a strict version of the constraint, and that is when $k = 3$. There are a couple of other "exceptions" (such as a model constrained by $\alpha A10$ or $\alpha B10$), but these are less interesting as they do not translate to the constraint (Z) or the relaxed version.

We wanted to confirm this *trend* of a model constrained by a higher k leads to, in most cases, a lower maximum LL, but instead for any $k \in \{1, 2, \ldots, 10\}$. Due to time constraints, we built two three-variable-models to analyse the trend using age, gender and the number of prior offences as variable inputs, and each model used either a strict version of the (A) constraint or (B) constraint. Figs 8 and 9 confirm that the 'trend' for all k values persist and the same violation at $k = 3$, also persists.

6.4.2 Violation of the k rule

While choosing a lower k value achieves results in a more accurate model, the consequence is that informativeness is (in most cases) compromised. The underlying Pareto improvement stems from the "violation" cases, where a planner can improve performance in two dimensions at no cost. We saw the unique case where a model constrained with $k = 3$ bins computes a higher maximum LL than does a model constrained with $k = 2$ bins, for any constraint Ω and type ω. Findings, as reported by Tables 4 – 9, indicate that $\hat{F}_3^{\omega}(\Omega) > \hat{F}_2^{\omega}(\Omega)$ for any constraint Ω and type ω. Thus, a model constrained by $\omega\Omega 3$ unambiguously improves performance in two dimensions (informativeness and accuracy) over a model constrained by $\omega\Omega 2$ for a given fairness rule.

Many traditional classifiers use this setting of a 2-bin condition (*low risk* and *high risk*), which is analogous to a binary classifier, even if they do not always set the bin boundary at 0.5. If instead, a scoring system of $k = 3$ was chosen (*low, medium and high risk*) over $k = 2$, the objective improvement in the payoff function can be equivocally expressed as $\hat{u}_3^{\omega}(\Omega) > \hat{u}_2^{\omega}(\Omega)$. Solution betas, $\hat{\beta}^k$, for all models (constrained and unconstrained) are reported in Tables 4 – 9.

6.4.3 Understanding the violation

To see why a 3-bin condition can be satisfied when a 2-bin condition is not, suppose $\alpha_{\mathrm{R}\delta}$ is the proportion of recidivists assigned to bin δ. We assume $R \in \{B, W\}$ where B denotes black recidivists, W denotes white recidivists, with $\delta \in \{1,2\}$ for a 2-bin system and $\delta \in \{1,2,3\}$ for a 3-bin system. The total number of recidivists in bin δ is denoted by $n_{R\delta}$. Thus, $\alpha_{\mathrm{B1}} + \alpha_{\mathrm{B2}} = 1$

and similarly, $\alpha_{W1} + \alpha_{W2} = 1$. It is easy to see that in a 2-bin scoring system, constraint (A) is met if, and only if, $\alpha_{B1} = \alpha_{W1}$:

$$\frac{(1)n_{B1} + (2)n_{B2}}{n_B} = \frac{(1)n_{W1} + (2)n_{W2}}{n_W} \qquad 1$$

$$(1)\alpha_{B1} + (2)(1 - \alpha_{B1}) = (1)\alpha_{W1} + (2)(1 - \alpha_{W1}) \qquad 2$$

$$\alpha_{B1} = \alpha_{W1} \qquad 3$$

Now suppose we use a 3-bin scoring system, $k = 3$, thus $\delta \in (1,2,3)$. Since $\alpha_{B1} + \alpha_{B2} + \alpha_{B3} = 1$ and $\alpha_{W1} + \alpha_{W2} + \alpha_{W3} = 1$, it is easy to see that constraint (A) is met if, and only if, $\alpha_{B3} - \alpha_{B1} = \alpha_{W3} - \alpha_{W1}$

$$\frac{(1)n_{B1} + (2)n_{B2} + (3)n_{B3}}{n_B} = \frac{(1)n_{W1} + (2)n_{W2} + (3)n_{W3}}{n_W} \qquad 4$$

$$(1)\alpha_{B1} + (2)(1 - \alpha_{B1} - \alpha_{B3}) + (3)\alpha_{B3} = (1)\alpha_{W1} + (2)(1 - \alpha_{W1} - \alpha_{W3}) + (3)\alpha_{W3} \qquad 5$$

$$\alpha_{B3} - \alpha_{B1} = \alpha_{W3} - \alpha_{W1} \qquad 6$$

Consider an arbitrary example where $n_B = n_W = 1000$. Suppose a particular model assigned recidivism probabilities as follows. Of the black recidivists, 250 were assigned probabilities below 0.3 and the other 750 probabilities above 0.7. Of the white recidivists, 125 were assigned probabilities below 0.3, another 625 were assigned probabilities above 0.7, and the remaining 250 were each assigned a probability between 0.55 and 0.6. This data is reported neatly in Table 11.

In this example, we meet the 3-bin version of (A) but not the 2-bin version of (A). Plugging in the data from the example above, when $k = 2$, we get:

$$\frac{(1)(250) + (2)(750)}{1000} \neq \frac{(1)(125) + (2)(875)}{1000}$$

The equation above deduces to $1.75 \neq 1.85$. Thus, balance for the positive class is not satisfied in a 2-bin scoring system with this data. Now, by plugging in the data when $k = 3$, we get:

$$\frac{(1)(250) + (2)(0) + (3)(750)}{1000} = \frac{(1)(125) + (2)(250) + (3)(625)}{1000}$$

The equation above deduces to $2.5 = 2.5$. Thus, balance for the positive class is satisfied in a 3-bin scoring system with this data.

7 Conclusion and further research

The Algorithmic Accountability Bill was approved in 2017 to address discrimination by regulating privatised RPIs. Following ProPublica's allegations, the challenge of creating a probabilistic classification that was equally fair to different groups captured the spotlight. Scholars and practitioners alike, have since stressed the mutual incompatibility between different notions of fairness. Despite the incompatibility, this dissertation motivates the discussion that we still *can improve efficiencies in prediction tools*. It attempts to clarify the gaps in the research, which also helps to establish the novelty of this work.

This research paper studied the sensitivity between the performance of the logistic tool and the chosen design parameters. I restricted attention to three dimensions and ran simulations that demonstrated the following: the relaxed version of a constraint allows significantly better accuracy than does a system constrained by the strict criterion, such that $\hat{F}_k^{\gamma}(\Omega) > \hat{F}_k^{\alpha}(\Omega)$ for any constraint Ω and bin number k; Figs 14 - 17 exhibit the probability distributions for strict versions of (Z), and by examining the divergence in these class distributions, the particular pattern we observe of a decreasing k with an increased maximum LL, is revealed; payoffs for the relaxed versions of the constraint can be expressed in the following functional form, $\hat{u}_9^{\gamma}(\Omega) > \hat{u}_3^{\gamma}(\Omega) > \hat{u}_2^{\gamma}(\Omega)$; A model constrained by $\gamma\Omega 9$ finds probability estimates that are not statistically significantly different to the estimates generated by the unconstrained model; for any choice set $k \in \{2, 3, 9, 10\}$ imposing relaxed versions of either (A) or (B) will unambiguously satisfy the relaxed version of (Z); experimental findings confirmed an improvement in performance in two dimensions (accuracy and informativeness) at a particular violation to the "k-rule" for a given fairness constraint. The corresponding objective improvement in the payoff function is expressed as $\hat{u}_3^{\omega}(\Omega) > \hat{u}_2^{\omega}(\Omega)$ for any constraint Ω and type ω.

7.1 Limitations

A key issue when conducting this research stems from the biases which might be implicitly embedded into the data. The assumption used throughout this paper is that observed recidivism is a suitable outcome measure for evaluating fairness. However, in the real world we only observe whether the criminal was *re-convicted* and not whether they did, in truth, *re-offend*. It is widely accepted knowledge that many criminal offenders are simply never identified. If a non-trivial fraction of individuals were identified as non-offenders but did in truth re-offend, then internal validity is compromised and the empirical assessments in this paper may be overstated.

7.2 Future work

A future goal of this research would be to build a model that is 'accurate out-of-sample' and one that serves a practical purpose outside of the criminal justice setting. As a robustness measure, I would employ a 'test' or 'hold out' data set. Since the prediction algorithm (formulated in this dissertation) had access to defendants' "outcomes" (recidivists and non-recidivists), the model might appear to perform well since it is evaluating data that it has already seen. In the future, to avoid the model being prone to a case of "unhelpful human data mining", I would randomly partition the data into "40% training, 40% imputation and 20% test data sets" (Kleinberg, Lakkaraju, et al. 2018). Such a procedure, of training the RPI on a partitioned dataset and then evaluating the performance on the 'hold-out data set, would not be a difficult task to carry out (but would be a welcomed addition to help validate findings).

If different judges sentenced identical populations at different rates, then we are presented with another case of implicit bias being embedded in the training data. Since we cannot faithfully observe the full breadth of unconstrained judicial preferences, we could have applied a common solution to this problem by imposing *independency constraints* (Calders, Kamiran, and Pechenizkiy 2009). Independency constraints are commonly used to offset or reverse the effect of bias by making the classifier's outcome independent of any *sensitive* variables, e.g., race, sex or religion. The fundamental problem with this method is that the more independent a model is of its sensitive variables, generally the lower the accuracy is of the classifier. For example, if ZIP codes were to exhibit a high degree of collinearity with ethnicity we might see a parallel drawn with the practice of *redlining*, e.g., where applicants were denied loans due to their residential area in which they inhabit.

If we can encourage practitioners to adopt the framework outlined in this dissertation, a fuller evaluation of the system should be carried out to encompass aspects of both validation and verification (and the two above examples could be a start!). A challenge would be building a refined architecture that scales well in higher dimensions.

7.3 Final thoughts

The initial question that inspired and drove this research was "how can algorithmic tools used in predictive analytics assist their human decision-makers in ameliorating potential discrimination and how should they be deployed to optimise the quality and fairness of social decision-making? Is there a trade-off between accuracy and fairness, and if so, how is this best managed?"

The body of literature in prediction algorithms and machine learning is so vast that it made for a very challenging research question. The more time I spent amongst the underlying literature and discussing the question with my supervisors, my domain understanding began to expand and as a result, the research became increasingly fascinating.

When writing this dissertation, I had the goal of developing a framework that would keep pace with innovations in contemporary literature. Applications of prediction techniques (as used in this paper) are as readily applicable in other domains such as: assessing the likelihood of a customer defaulting on a loan; identifying risk factors for different diseases; predicting the likelihood of voting for a particular candidate in an election; predicting which teacher will have the highest value-added to an institution (Rockoff et al. 2011); and in targeting health inspections (Kang et al. 2013). These results illustrate just a few examples of what we could even conceive to be predictable – the opportunities are almost endless.

The core of this research challenge was to present 'compelling enough' motivation for adopting the framework built in this paper, that was both feasible and pragmatic. "Pragmatism adopts an engineering approach to research - it values practical knowledge over abstract knowledge, and uses whatever methods are appropriate to obtain it" (Easterbrook et al., 2008, page 292). By combining the literature review findings with practical experimentation, we have shown in a pragmatic sense, how the applicability of a harm-reduction framework can be used in predictive policing. Being able to extend the capabilities of the framework to heterogeneous domains (or more dimensions) would make for an exciting future research challenge. In the meantime, while we wait for theory to progress, there are practical solutions that can be applied to promote improved efficiencies in our social welfare systems.

8 Bibliography

Angwin, Julia and Jeff Larson. 2016. “Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say.” *ProPublica*.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. “Machine Bias: There’s Software Used across the Country to Predict Future Criminals. And It’s Biased against Blacks.” *ProPublica*.

Armitage, P. and G. Berry. 1997. *Statistical Methods in Medical Research.* Vol. 53.

Brennan, Tim, William Dieterich, and Beate Ehret. 2009. “Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System.” *Criminal Justice and Behavior*.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy. 2009. “Building Classifiers with Independency Constraints.” in *ICDM Workshops 2009 - IEEE International Conference on Data Mining*.

Chouldechova, Alexandra. 2017. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data*.

Collings, Stanley, J. Aitchison, and I. R. Dunsmore. 1981. “Statistical Prediction Analysis.” *The Mathematical Gazette*.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, and Sharad Goel. 2016. “A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks . It’ s Actually Not That Clear.” *The Washington Post*.

Courtland, Rachel. 2018. “Bias Detectives: The Researchers Striving to Make Algorithms Fair News-Feature.” *Nature*.

Cowgill, Bo and Catherine E. Tucker. 2019. “Economics, Fairness and Algorithmic Bias.” *SSRN Electronic Journal*.

Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. “COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.” *Northpointe Inc*.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. “Fairness through Awareness.” in *ITCS 2012 - Innovations in Theoretical Computer Science Conference*.

Easterbrook, Steve, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. “Selecting Empirical Methods for Software Engineering Research Guide to Advanced Empirical Software Engineering.” *Guide to Advanced Empirical Software Engineering*.

FBI. 2014. “Federal Bureau of Investigation, Uniform Crime Report.” *Federal Bureau of Investigation, Uniform Crime Report*.

Friedman, Jerome H. 1997. “On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality.” *Data Mining and Knowledge Discovery*.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. “Equality of Opportunity in Supervised Learning.” in *Advances in Neural Information Processing Systems*.

Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. “Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews.” in *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Kirchner, Lauren. 2017. “New York City Moves to Create Accountability for Algorithms.” *ProPublica*.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. “Human Decisions and Machine Predictions.” *Quarterly Journal of Economics*.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. “Prediction Policy Problems.” in *American Economic Review*.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. *Algorithmic Fairness*. Vol. 108.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” in *Leibniz International Proceedings in Informatics, LIPIcs*.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. “How We Analyzed the COMPAS Recidivism Algorithm.” *ProPublica*.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. “On Fairness and Calibration.” in *Advances in Neural Information Processing Systems*.

Rockoff, J. E., B. A. Jacob, T. J. Kane, and D. O. Staiger. 2011. “Can You Recognize an Effective Teacher When You Recruit One?” *Education Finance and Policy*.

Singh, Jay P. 2013. “Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer.” *Behavioral Sciences and the Law*.

9 Figures

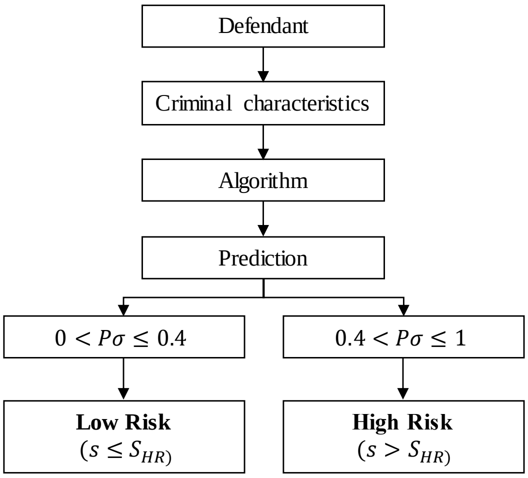

Figure 1: Algorithmic pipeline. Illustrates the process of mapping probability estimates to scores.

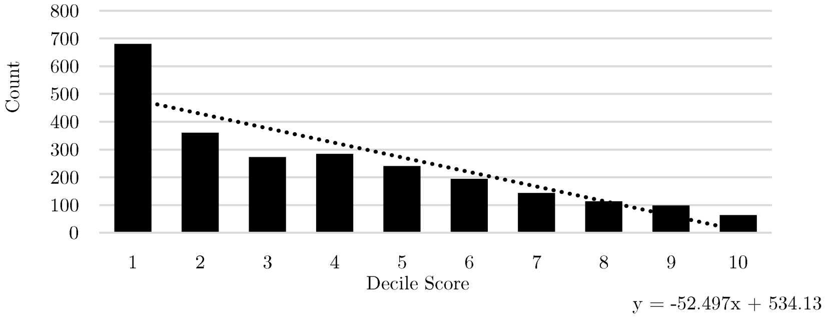

Figure 2: COMPAS risk score distribution for Caucasians

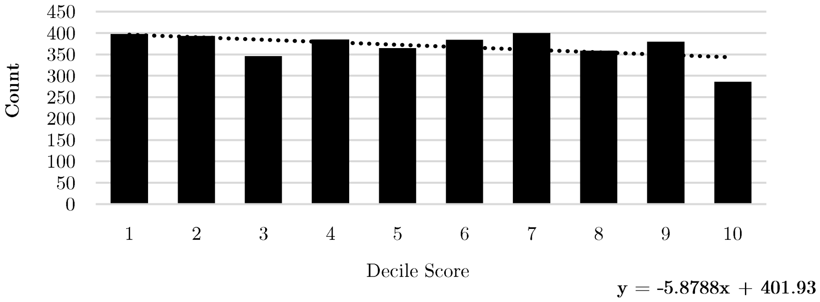

Figure 3: COMPAS risk score distribution for African-Americans

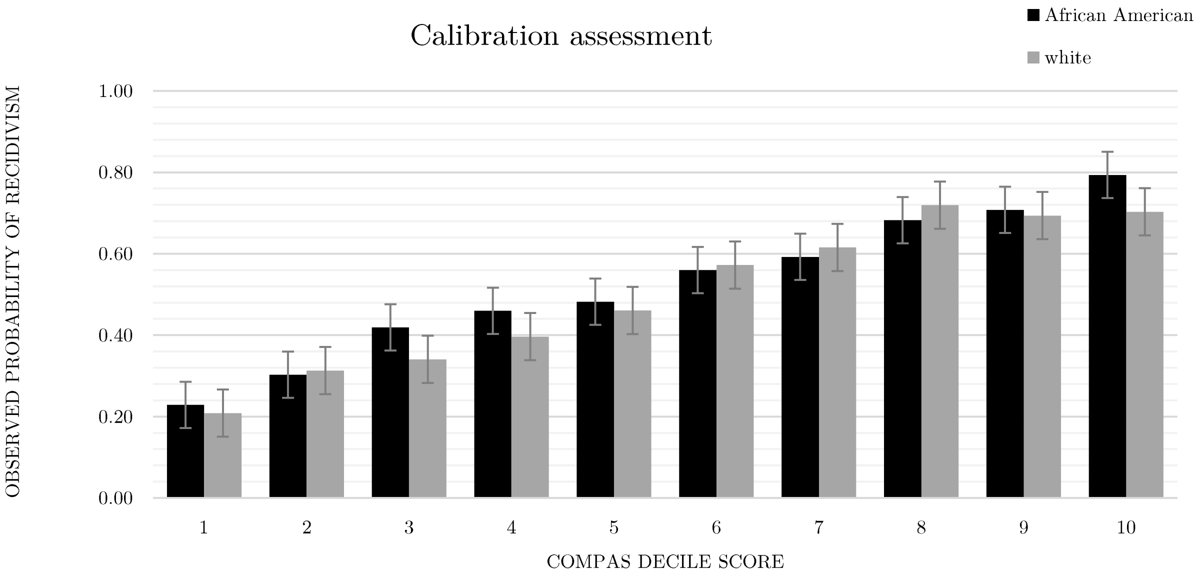

Figure 4: Calibration assessment of COMPAS. Figure illustrates that the tool is in fact *well-calibrated* as score differentials in each "bin" are non-significant. Note that standard error bars overlap between African-Americans and Caucasians at any given score.

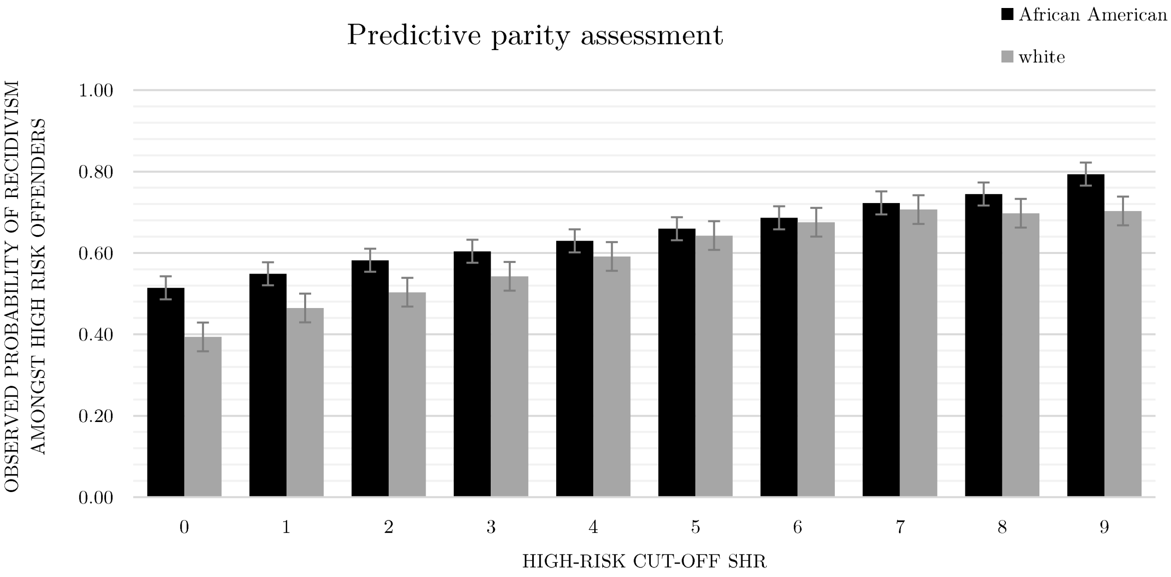

Figure 5: Predictive parity assessment of COMPAS. Figure illustrates that the tool satisfies *Predictive parity*. Note that standard error bars overlap between African-Americans and Caucasians at all thresholds with the exception of $S_\sigma^k = 0$ and $S_\sigma^k = 9$. This might be due to the *degree of error* a decision-maker chooses to compare the risk scores between both race groups with. Differentials are non-significant for all other high risk cut-off thresholds.

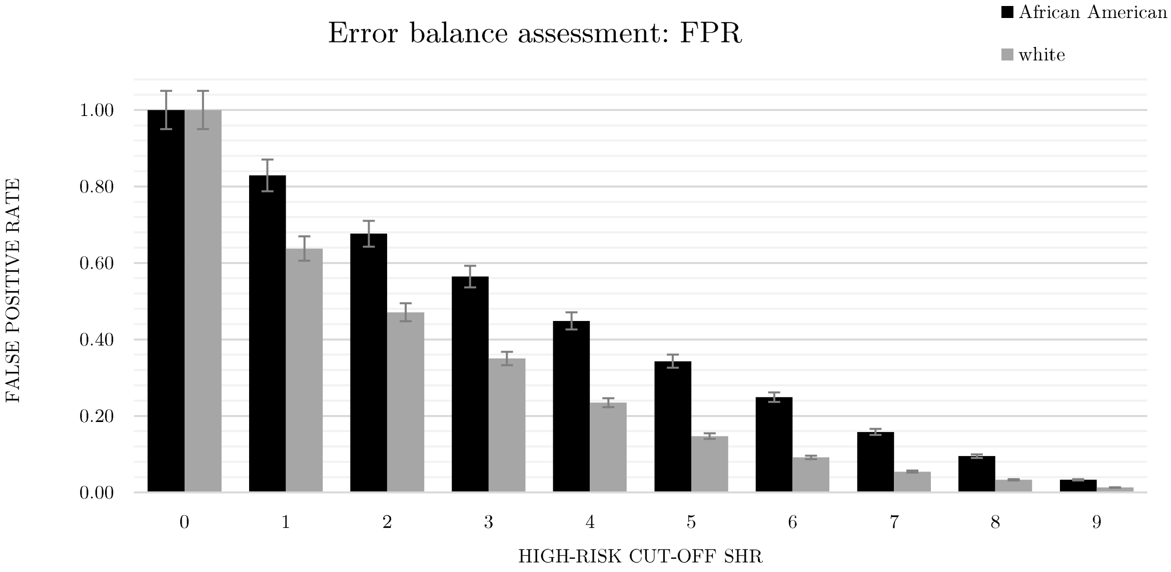

Figure 6: False positive rate assessment of the COMPAS risk tool. Figure illustrates that P(S > S_{HR} | Y = 0, r) for the values of the high risk cut-off thresholds. Figure illustrates that false positive rates between group membership are significantly different across all scores where S> $S_{HR.}$

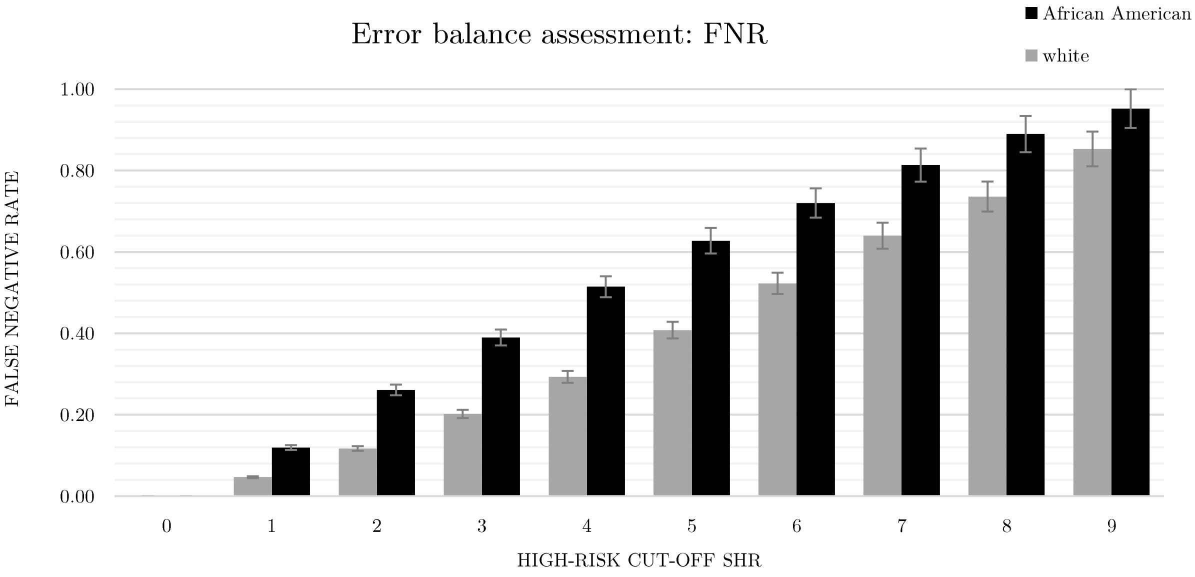

Figure 7: False negative rate assessment of the COMPAS risk tool. Bars represent the expression: P(S ≤ S_{HR} | Y = 1, r) for scores from the high risk cut-off thresholds. Figure illustrates that false negative rates between group membership are significantly different across all scores where S≤ $S_{HR.}$

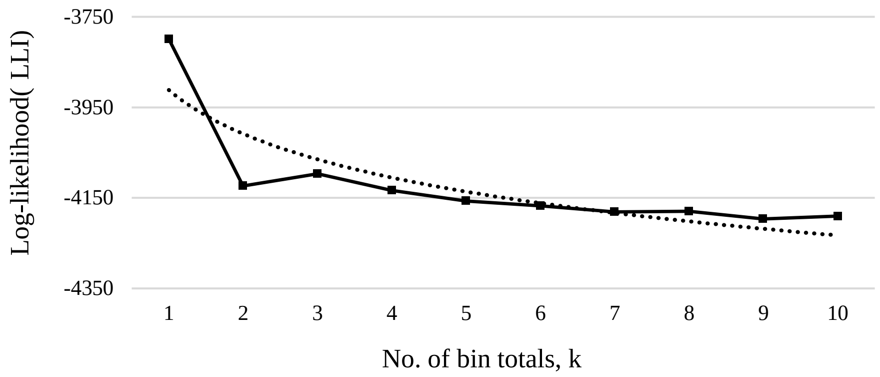

Figure 8: Sensitivity of accuracy to bin totals – negative class constraint

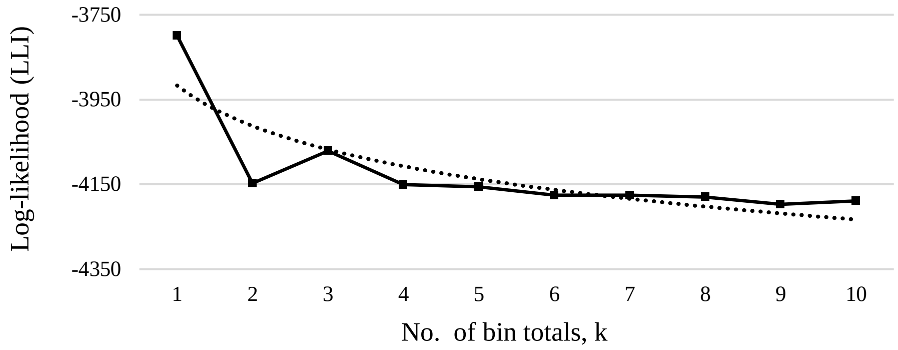

Figure 9: Sensitivity of accuracy to bin totals – positive class constraint.

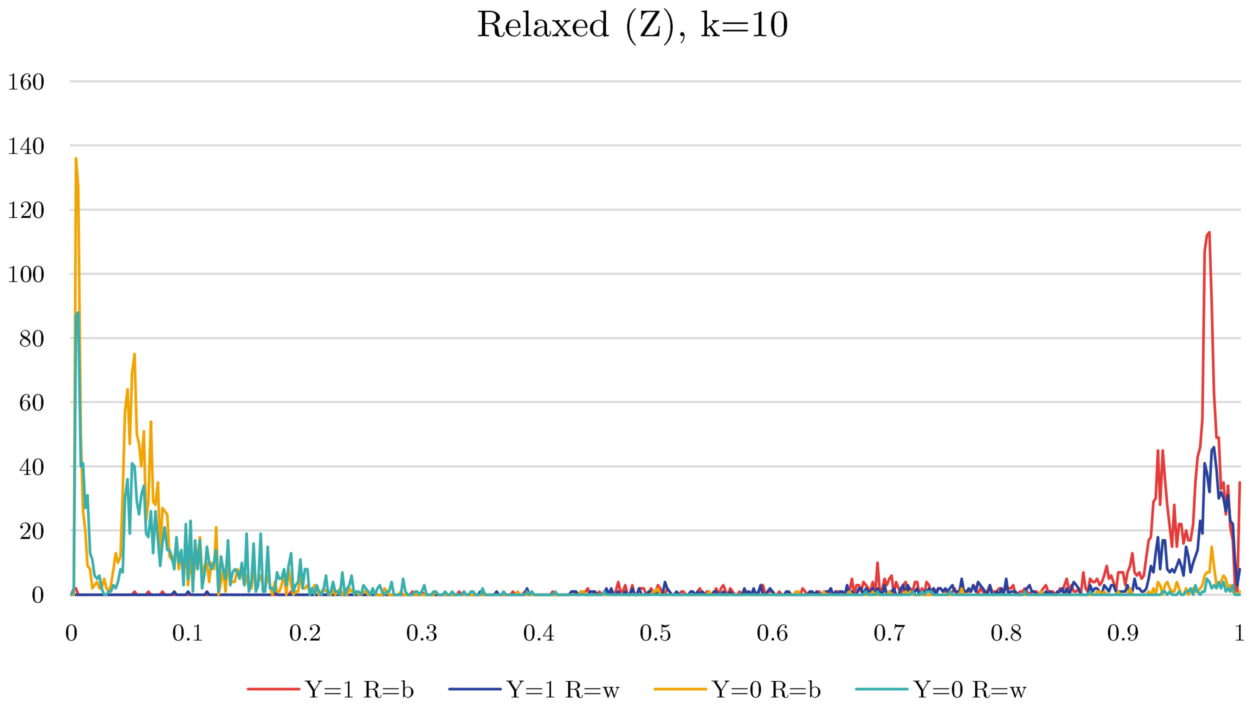

Figure 10: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a relaxed version of (Z) using a 10-bin scoring system.

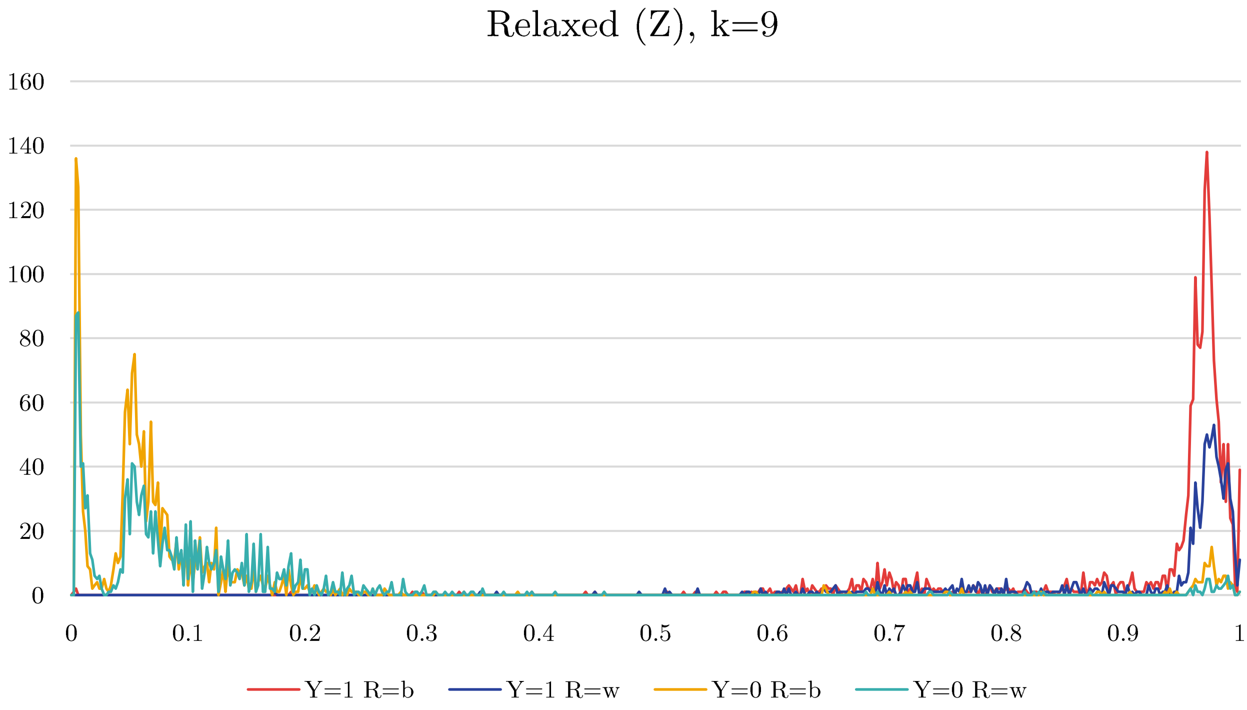

Figure 11: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a relaxed version of (Z) using a 9-bin scoring system.

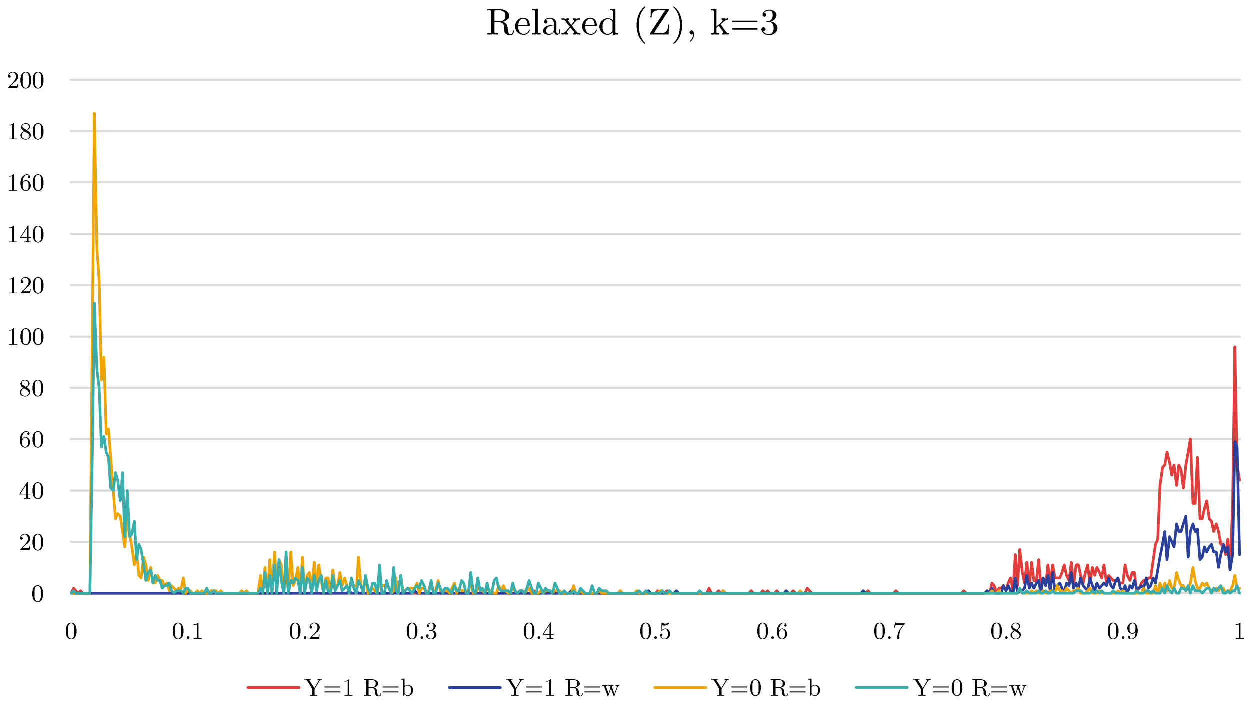

Figure 12: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a relaxed version of (Z) using a 3-bin scoring system.

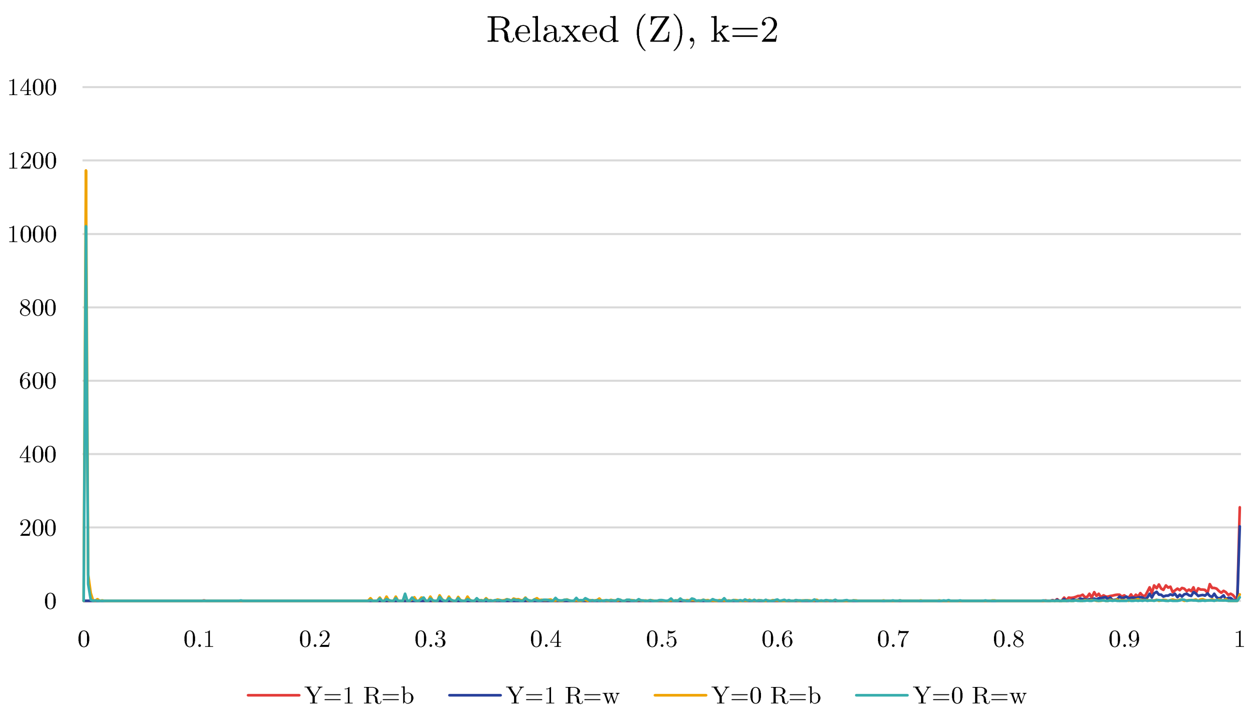

Figure 13: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a relaxed version of (Z) using a 2-bin scoring system.

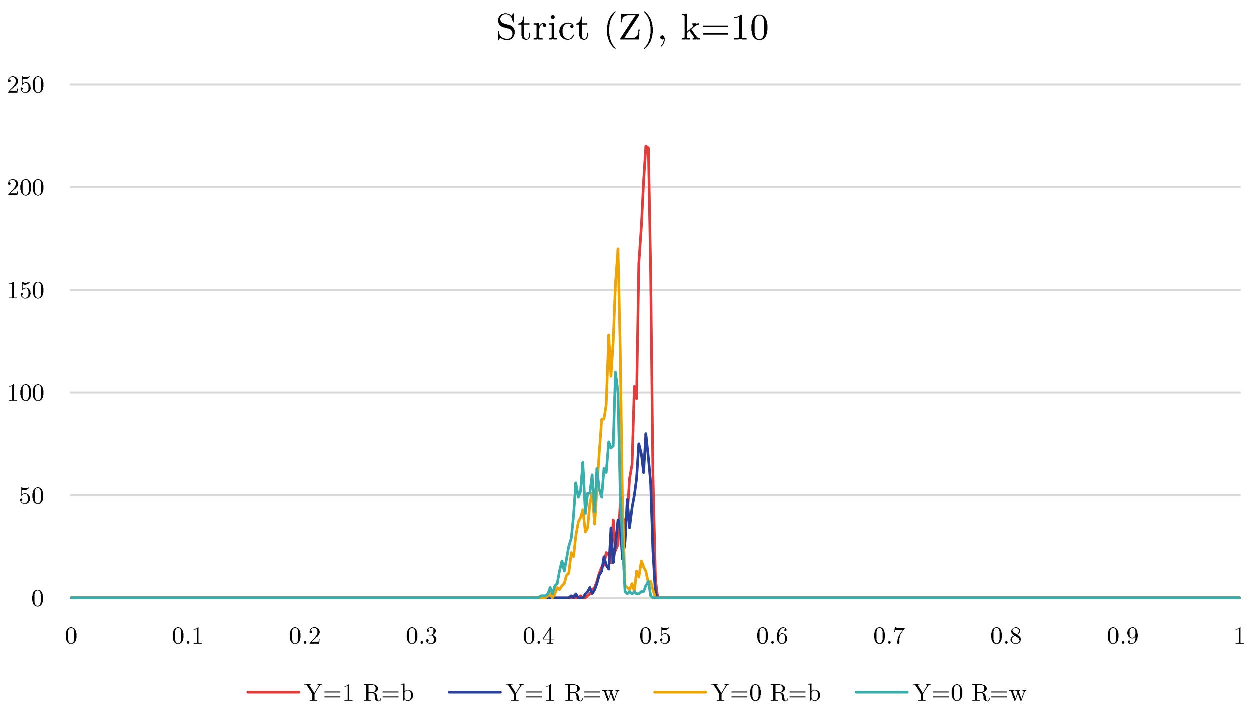

Figure 14: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a strict version of (Z) using a 10-bin scoring system.

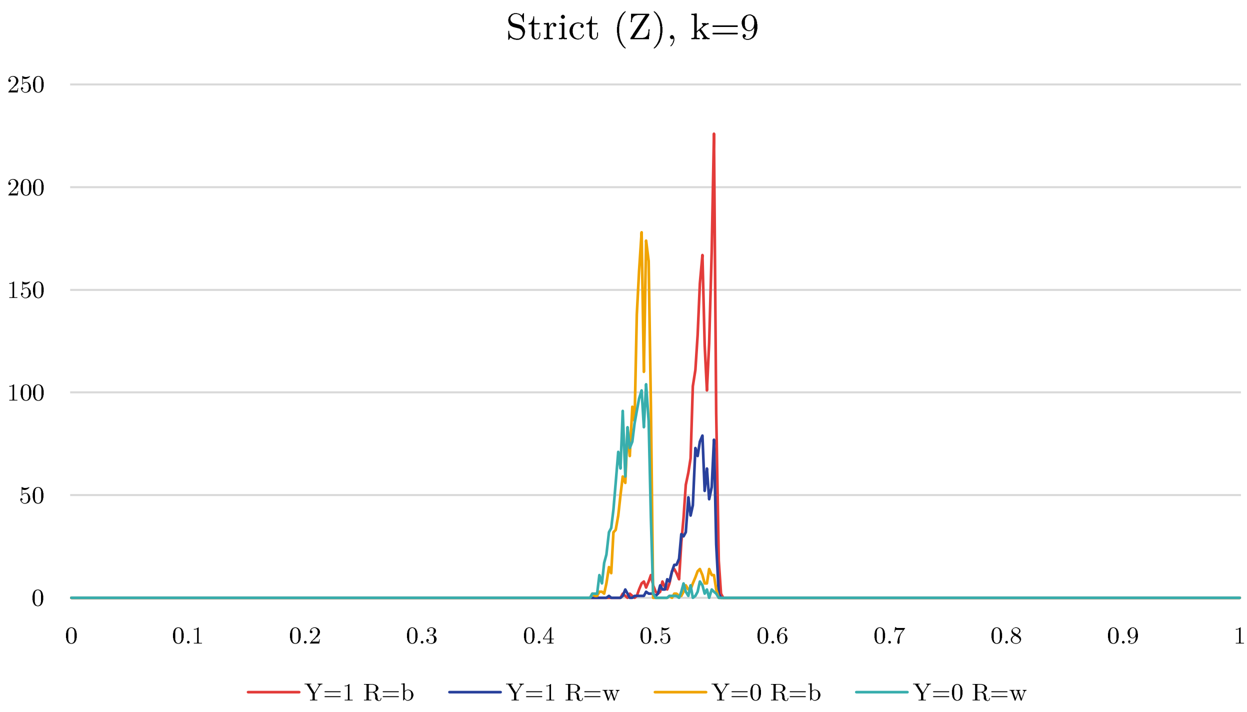

Figure 15: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a strict version of (Z) using a 9-bin scoring system.

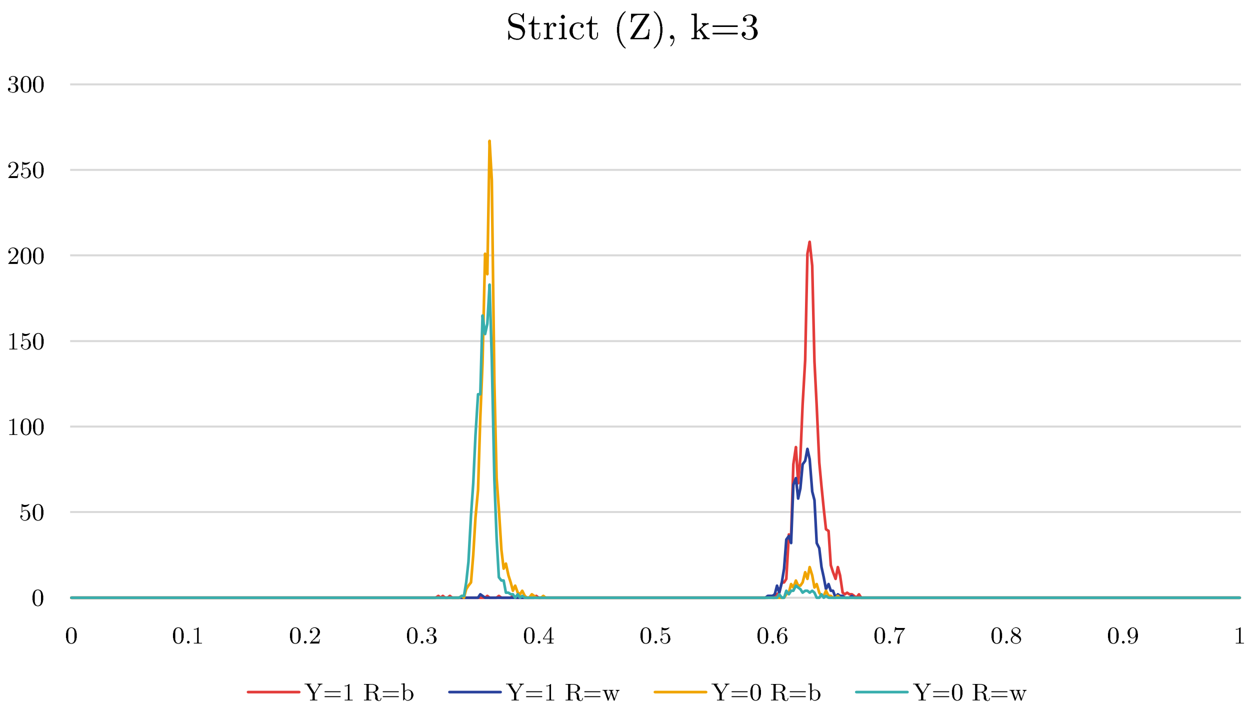

Figure 16: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a strict version of (Z) using a 3-bin scoring system.

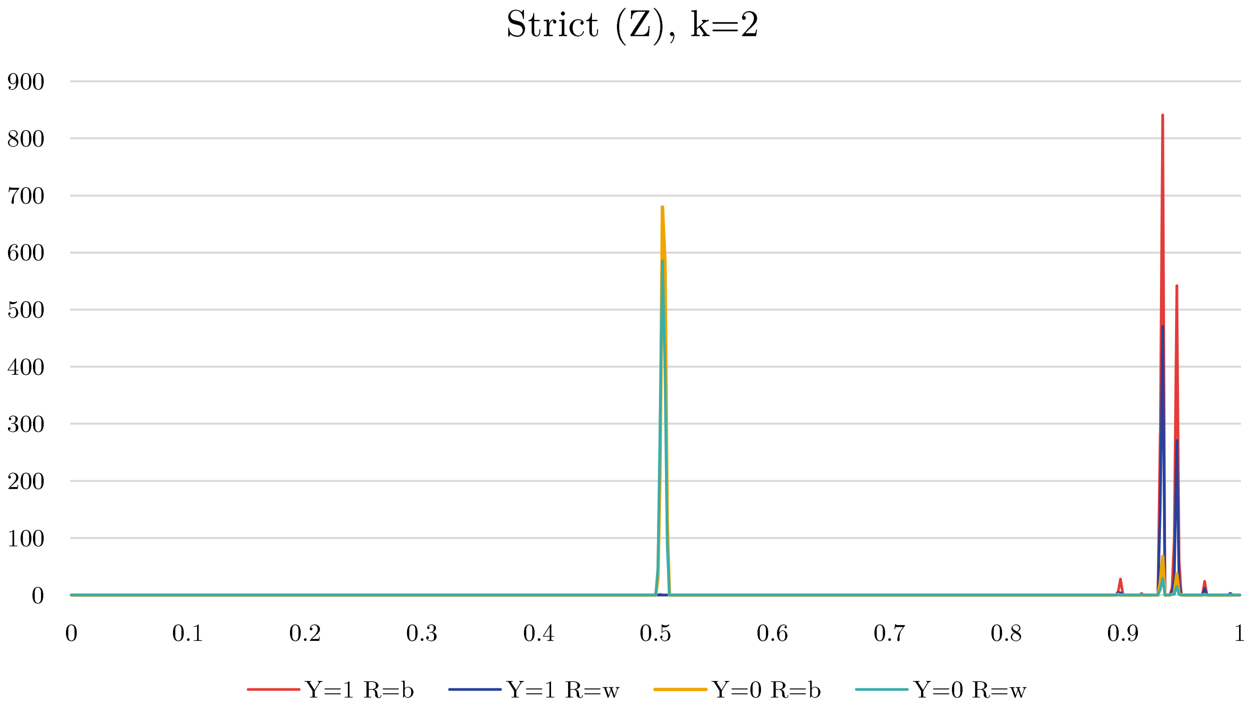

Figure 17: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the model constrained by a strict version of (Z) using a 2-bin scoring system.

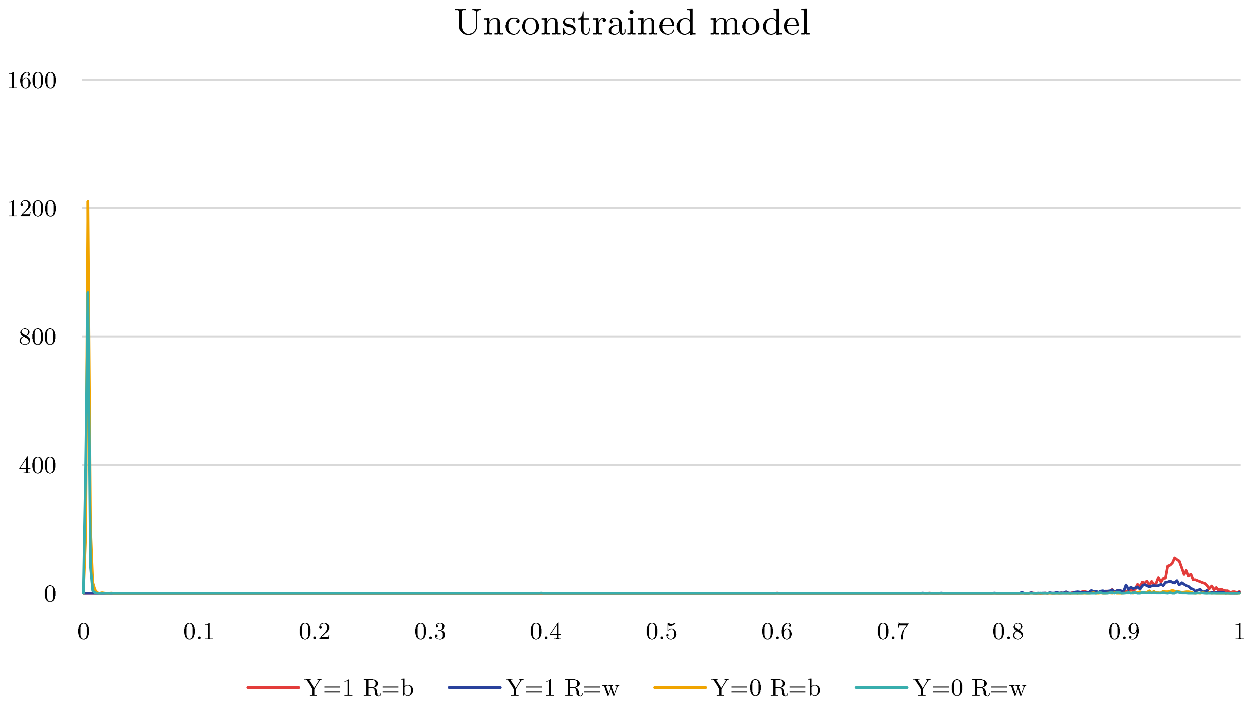

Figure 18: Distribution of probability estimates for the positive (Y=1) and negative (Y=0) classes for the Unconstrained model

10 Tables

Table 1a: Summary statistics for African-Americans (r=b) and Caucasians (r=w)

Variable	mean	mean (r=b)	mean (r=w)	T-test
Age	34.73 (11.91)	32.74 (10.96)	37.73 (12.75)	15.91
Ave no. prior offences	3.70 (5.03)	5.16 (5.67)	2.43 (3.79)	15.49
Risk score	4.72 (2.85)	5.79 (2.86)	3.78 (2.59)	23.29
Violent risk score	3.82 (2.52)	4.43 (2.59)	2.95 (2.20)	23.52

Table 2b: Charge degree for African-Americans (r=b) and Caucasians (r=w)

Variable	total	r=b	r=w
Felonies	1006	676	330
Misdemeanors & other	2055	1360	695

Table 3: Summary of risk scores by Recidivism, Race and Sex

Label	Total	Average risk score	% of subset	% of total	Average risk score ratio of subset	Average risk score ratio (of total)
r=b	3696	5.37	-	60.10%	-	1.14
r=w	2454	3.74	-	39.90%	-	0.79
s=f	1219	4.36	-	19.82%	-	0.92
s=m	4931	4.81	-	80.18%	-	1.02
Y=1	2867	5.79	-	46.62%	-	1.23
Y=0	3283	3.78	-	53.38%	-	0.80

Label	Total	Average risk score	$= (r\|Y)$	% of total	Average risk score ratio $\in (Y)$	Average risk score ratio (of total)
Y=1, r=b	1901	6.29	66.31%	30.91%	1.09	1.33
Y=1, r=w	966	4.82	33.69%	15.71%	0.83	1.02
Y=0, r=b	1795	4.40	54.68%	29.19%	1.16	0.93
Y=0, r=w	1488	3.03	45.32%	24.20%	0.80	0.64

Label	Total	Average risk score	$= (s\|Y)$	% of total	Average risk score ratio $\in (Y)$	Average risk score ratio (of total)
Y=1, s=f	446	5.49	15.56%	7.25%	0.95	1.16
Y=1, s=m	2421	5.85	84.44%	39.37%	1.01	1.24
Y=0, s=f	773	3.70	23.55%	12.57%	0.98	0.79
Y=0, s=m	2510	3.80	76.45%	40.81%	1.01	0.81

Label	Total	Average risk score	$= (s\|r, Y)$		$= (r\|s, Y)$	
Y=1, r=b, s=f	247	5.95	12.99%		55.38%	
Y=1, r=b, s=m	1654	6.34	87.01%		68.32%	
Y=1, r=w, s=f	199	4.92	20.60%		44.62%	
Y=1, r=w, s=m	767	4.79	79.40%		31.68%	
Y=0, r=b, s=f	405	4.03	22.56%		52.39%	
Y=0, r=b, s=m	1390	4.50	77.44%		55.38%	
Y=0, r=w, s=f	368	3.34	24.73%		47.61%	
Y=0, r=w, s=m	1120	2.93	75.27%		44.62%	

Table 4: Solution values and maximum log-likelihood for unconstrained model

Var (table 11	*Value*
B0	-5.0318
B1 - # Priors	0.0650
B2 - Age	-0.0392
B3 - Gender	-0.0262
(F1)	9.8124
(F2)	8.6303
(F3)	7.9830
(F6)	12.787
(F7)	41.790
(M1)	8.3008
(M2)	8.4310
(MO3)	9.0150
(CO3)	5.6840
$\boldsymbol{\hat{F}(\beta, X)}$	**-722.65**

Table 5: Solution values and maximum log-likelihood using a “strict” version of (Z)

Var	*k=2*	*k=3*	*k=9*	*k=10*
B0	0.0359	-0.5449	0.0453	-0.0234
B1 - # Priors	-0.0002	0.0062	0.0013	0.0010
B2 - Age	0.0077	-0.0217	-0.0290	-0.0055
B3 - Gender	-0.0004	-0.0017	-0.0032	-0.0046
(F1)	2.1378	1.0932	0.0054	0.0555
(F2)	2.7719	1.1372	0.1713	0.1069
(F3)	2.5940	1.0596	0.2220	0.0909
(F6)	2.3584	-0.2045	0.0439	0.1009
(F7)	4.7754	-0.0375	0.0041	0.1335
(M1)	2.6131	1.1069	0.2208	0.1011
(M2)	2.8330	1.1203	0.1812	0.0976
(MO3)	3.4017	1.1321	0.0410	0.1152
(CO3)	-0.0136	0.1292	0.0862	-0.1168
$\boldsymbol{\hat{F}_k^{\alpha}(Z)}$	**-2896.88**	**-2878.40**	**-3962.07**	**-4076.82**

Table 6: Solution values and maximum log-likelihood using a “strict” version of (A)

Var	*k=2*	*k=3*	*k=9*	*k=10*
B0	0.0359	-0.5560	0.0502	0.0812
B1 - # Priors	-0.0002	0.0061	0.0014	0.0021
B2 - Age	0.0077	-0.0117	-0.0457	-0.0778
B3 - Gender	-0.0004	-0.0017	-0.0042	-0.0069
(F1)	2.1378	1.1678	0.0096	0.0243
(F2)	2.7719	1.1436	0.2069	0.0419
(F3)	2.5940	1.1174	0.2387	1.1171
(F6)	2.3584	-0.1459	0.0732	-0.8208
(F7)	4.7754	1.1733	0.0070	0.0829
(M1)	2.6131	1.1211	0.2397	0.1680
(M2)	2.8330	1.1356	0.2474	0.0397
(MO3)	3.4017	1.2002	0.0660	-0.0095
(CO3)	-0.0136	0.5931	0.0988	-0.4552
$\boldsymbol{\widehat{F}_k^{\alpha}(A)}$	**-2896.88**	**-2849.04**	**-3905.39**	**-3793.09**

Table 7: Solution values and maximum log-likelihood using a “strict” version of (B).

Var	*k=2*	*k=3*	*k=9*	*k=10*
B0	0.0359	-0.5449	0.0507	0.0041
B1 - # Priors	-0.0002	0.0062	0.0023	0.0002
B2 - Age	0.0077	-0.0217	-0.0269	-0.0009
B3 - Gender	-0.0004	-0.0017	-0.0036	0.0000
(F1)	2.1378	1.0932	-0.0690	0.0452
(F2)	2.7719	1.1372	0.2038	0.0062
(F3)	2.5940	1.0596	0.2462	0.3975
(F6)	2.3584	-0.2045	0.0878	-0.6948
(F7)	4.7754	-0.0375	0.3402	1.1303
(M1)	2.6131	1.1069	0.2441	0.3994
(M2)	2.8330	1.1203	0.2411	0.3771
(MO3)	3.4017	1.1321	0.0917	0.0243
(CO3)	-0.0136	0.1292	0.2131	0.3046
$\boldsymbol{\widehat{F}_k^{\alpha}(B)}$	**-2896.88**	**-2878.40**	**-3913.47**	**-3834.17**

Table 8: Solution values and maximum log-likelihood using a “relaxed” version of (Z).

Var	*k=2*	*k=3*	*k=9*	*k=10*
B0	-8.7718	-4.7098	-4.3124	-4.1170
B1 - # Priors	0.0791	0.0355	-0.0014	-0.0554
B2 - Age	6.9274	2.4306	-2.0171	-2.7171
B3 - Gender	0.0367	0.0311	0.0317	0.0447
(F1)	10.0974	10.5531	19.2753	23.6664
(F2)	9.8509	5.3681	4.8435	5.0012
(F3)	9.7109	5.4825	6.8125	6.5560
(F6)	47.4068	-2.4421	2.1185	-3.3132
(F7)	15.1000	2.7605	33.6485	16.3794
(M1)	10.4637	6.6904	6.1655	6.2126
(M2)	9.8060	6.6111	7.4651	6.5008
(MO3)	9.6348	3.6911	5.8234	4.3934
(CO3)	5.0617	-4.8416	-4.5306	19.5780
$\boldsymbol{\hat{F}_k^{\gamma}(Z)}$	**-1321.98**	**-1145.99**	**-976.69**	**-1161.90**

Table 9: Solution values and maximum log-likelihood using a “relaxed” version of (A).

Var	*k=2*	*k=3*	*k=9*	*k=10*
B0	-8.7718	-4.7098	-4.3124	-4.1170
B1 - # Priors	0.0791	0.0355	-0.0014	-0.0554
B2 - Age	6.9274	2.4306	-2.0171	-2.7171
B3 - Gender	0.0367	0.0311	0.0317	0.0447
(F1)	10.0974	10.5531	19.2753	23.6664
(F2)	9.8509	5.3681	4.8435	5.0012
(F3)	9.7109	5.4825	6.8125	6.5560
(F6)	47.4068	-2.4421	2.1185	-3.3132
(F7)	15.1000	2.7605	33.6485	16.3794
(M1)	10.4637	6.6904	6.1655	6.2126
(M2)	9.8060	6.6111	7.4651	6.5008
(MO3)	9.6348	3.6911	5.8234	4.3934
(CO3)	5.0617	-4.8416	-4.5306	19.5780
$\boldsymbol{\hat{F}_k^{\gamma}(A)}$	**-1321.98**	**-1145.99**	**-976.69**	**-1161.90**

Table 10: Solution values and maximum log-likelihood using a “relaxed” version of (B).

Var	*k=2*	*k=3*	*k=9*	*k=10*
B0	-8.7718	-4.7098	-4.3124	-4.1170
B1 - # Priors	0.0791	0.0355	-0.0014	-0.0554
B2 - Age	6.9274	2.4306	-2.0171	-2.7171
B3 - Gender	0.0367	0.0311	0.0317	0.0447
(F1)	10.0974	10.5531	19.2753	23.6664
(F2)	9.8509	5.3681	4.8435	5.0012
(F3)	9.7109	5.4825	6.8125	6.5560
(F6)	47.4068	-2.4421	2.1185	-3.3132
(F7)	15.1000	2.7605	33.6485	16.3794
(M1)	10.4637	6.6904	6.1655	6.2126
(M2)	9.8060	6.6111	7.4651	6.5008
(MO3)	9.6348	3.6911	5.8234	4.3934
(CO3)	5.0617	-4.8416	-4.5306	19.5780
$\widehat{\boldsymbol{F}}_{\boldsymbol{k}}^{\boldsymbol{\gamma}}(\boldsymbol{B})$	-1321.98	-1145.99	**-976.69**	**-1161.90**

Table 11: t-Test results for a relaxed version of (Z)

		t-Test relaxed (Z)		
Bins	$k = 2$	$k = 3$	$k = 9$	$k = 10$
T-stat (A)	1.0598	0.7213	0.4023	1.5259
T-stat (B)	0.3772	1.9265	1.4367	1.9597

Table 12: Variables for arbitrary examples 1 and 2

k=3, Y=1			
$\boldsymbol{P_\sigma}$	$\boldsymbol{n_B}$	$\boldsymbol{n_W}$	$\boldsymbol{\delta}$
$0 < P_\sigma < 0.35$	250	125	1
$0.5 \leq P_\sigma < 0.7$	0	250	2
$0.7 \leq P_\sigma < 1$	750	625	3

k=2 , Y=1			
$\boldsymbol{P_\sigma}$	$\boldsymbol{n_B}$	$\boldsymbol{n_W}$	$\boldsymbol{\delta}$
$0 < P_\sigma < 0.5$	250	125	1
$0.5 \leq P_\sigma < 1$	750	875	2

Table 13: Average probability estimates for relaxed and strict versions of (Z)

Average probabilities, k=10, strict (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.4838	0.4789	0.4566	0.4495
Max	0.5000		
Min	0.4004		

Average probabilities, k=9, strict (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.5368	0.5336	0.4866	0.4801
Max	0.5554		
Min	0.4444		

Average probabilities, k=3, strict (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.6298	0.6249	0.3767	0.3632
Max	0.6724		
Min	0.3132		

Average probabilities, k=10, relaxed (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.9145	0.9111	0.1190	0.1052
Max	1.0000		
Min	0.0015		

Average probabilities, k=9, relaxed (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.9313	0.9274	0.1008	0.0735
Max	1.0000		
Min	0.0003		

Average probabilities, k=2, strict (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.9369	0.9372	0.5380	0.5220
Max	0.9918		
Min	0.5000		

Average probabilities, k=3, relaxed (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.9294	0.9388	0.1456	0.1273
Max	0.9997		
Min	0.0002		

Average probabilities, k=2, relaxed (Z)

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.9421	0.9483	0.1554	0.1399
Max	1.0000		
Min	0.0003		

Average probabilities, unconstrained

Y=1, r=b	Y=1, r=w	Y=0, r=b	Y=0, r=w
0.9401	0.9246	0.0732	0.0386
Max	1.0000		
Min	0.0008		

Table 14: t-Test: probabilities of unconstrained vs. k=9, relaxed (Z) models (assuming unequal variances)

	k=9, relaxed (Z)	*unconstrained*
Mean	0.480734134	0.466505272
Variance	0.204192059	0.217260709
Observations	6150	6150
Hypothesized Mean Difference	0	
df	12286	
t Stat	1.718831377	
P(T<=t) one-tail	0.042835136	
t Critical one-tail	1.644977661	
P(T<=t) two-tail	0.085670271	
t Critical two-tail	1.960157091	