How Data Scientists Help Regulators and Banks Ensure Fairness when Implementing Machine Learning and Artificial Intelligence Models

Nicholas Schmidt[1] Bernard Siskin[2] Syeed Mansur[3]

February 1, 2018

Introduction

Nearly all major lending institutions are taking steps to implement machine learning and artificial intelligence models to better measure marketing response rates and creditworthiness, assess the likelihood of fraud, and quantify other risks and opportunities. This is no surprise: the increased effectiveness of these techniques over traditional clustering and GLM models is often staggering. One bank with which we work has seen about a 20% average increase in statistical lift as a result of moving from logistic regressions to gradient boosting machines.

However, implementation of machine learning has proved difficult because banks are highly regulated and face substantial challenges in complying with model governance standards and anti-discrimination laws. A distinct challenge has been that the guidelines for compliance have not kept up with the breakneck speed of technological change in the ML/AI universe. As a result, banks do not have certainty that they are in compliance with some regulatory requirements when utilizing machine learning models. In particular, we see a significant amount of uncertainty revolving around compliance with fair lending requirements of anti-discrimination law.[4]

This paper provides a high-level overview of these issues, and also outlines how we advise banks to comply with fair lending requirements when implementing machine learning. We describe the current focus of regulators, explain how discrimination is typically measured, and relate how we see fair lending compliance being implemented in the coming years.[5]

Regulatory Concerns with Machine Learning and AI Modeling

Banking regulators have expressed many of the same fairness concerns about machine learning that have been expressed in the media, and which form at least some of the bases for this conference. In particular, regulators worry about the limited degree to which machine learning processes provide explanations for their decisions. They ask, "If we do not know how the decision was made, how can we be sure that it was

[1] Partner, BLDS, LLC. Email: nschmidt@bldsllc.com
[2] Founder, BLDS, LLC. Email: bsiskin@bldsllc.com
[3] CEO, Sentrana, Inc. and DeepCortex.ai. Email: syeed.mansur@deepcortex.ai
[4] In this paper we address the issue of "fairness" within the context of U.S. regulatory precedent and practice. Certainly, other definitions of fairness and approaches to accomplishing fairness exist, most of which we expect to be compatible with the approach outlined herein.
[5] This paper focuses on banking because we have observed that banks are more interested than any other type of institution with which we work in maintaining anti-discriminatory compliance, while implementing machine learning. However, nearly identical questions are being asked by businesses that are not as regulated as banking, but are concerned with how their employment practices comply with anti-discrimination laws.

not based on the customer's protected class status (e.g., race or gender)?" The specific concern is that, even if a model does not explicitly include protected class status as a variable, some combination of variables becomes a proxy for protected class status. For example, if a model includes a customer's "likes" from a social media site, then it is easy to imagine that some combination of those "likes" can become a near perfect proxy for gender for at least a sub-group of customers.

Regulators have additional concerns when the outcome of a model is correlated with a protected class, even if the model does not use protected class status as an explanatory feature. They ask, "Is this the least discriminatory model possible?" In other words, can the relationship with protected class status be broken, or at least diminished, while still maintaining model quality.

On the other hand, there is enthusiasm among regulators for these models because of their increased accuracy. One of the goals of regulators is to make sure "the right customers are getting the right product" given the customer's credit profile. For instance, regulators are deeply concerned with customers being offered excessive or "predatory" loans that they will almost certainly be unable to repay. Because of the relative creditworthiness of different races, decisions to make loans to more creditworthy applicants will almost certainly have a negative average impact on some protected classes. However, regulators will generally be satisfied that those differences are reasonable, if the bank can show that this effect is due to protecting people (including protected class members) from excessive loans.

Anti-Discrimination Regulation and Types of Discrimination Recognized by Courts

Courts and regulators typically recognize two types of discrimination. The first is "Disparate Treatment", and the second is "Disparate Impact". Disparate treatment exists when a decision is made that explicitly takes into account protected class status, resulting in different outcomes for class members who are otherwise "similarly situated" to non-class members. A model causes disparate treatment when either protected class status is directly included in the model (e.g., a categorical feature indicating female/male), or a feature is included that is essentially a proxy for protected class status. A decision driven by disparate treatment will almost always be illegal in employment, housing, and banking, according to U.S. law.

The second type of discrimination, disparate impact, results from the inclusion of features that are disproportionately and negatively correlated with protected class status. In such a case, the model will have a disparate impact. Importantly, a finding of disparate impact does not mean that a model cannot be used. Instead, such a model is not considered legally discriminatory if it meets two criteria: (1) the use of the model is business justified (i.e., it is a valid predictor of the desired business outcome, such as job performance), and (2) there are no less discriminatory, but still business-justified, alternative model available.

For example, banking default models usually include some measure of creditworthiness, such as a FICO score. For most products African-Americans and Hispanics have, on average, lower FICO scores than whites. Because credit models predict that lower FICO scores lead to higher likelihood of default, African-Americans and Hispanics will receive fewer offers of credit. This means that a model that is largely based on FICO scores will likely have a disparate impact on African-Americans and Hispanics. However, because FICO scores have generally been shown to be valid and robust predictors of default, this disparate impact on Hispanics and African-Americans is likely to be legal because the use of the

model is business justified (i.e., it is a valid predictor of the outcome). The only question then is whether an equally business justified model with less disparate impact can be found.

In traditional GLMs, the process of finding less discriminatory models can be done relatively easily, as one can readily calculate each variable's discriminatory impact on the outcome, and then make changes to the model to lessen that impact. But no such simple process exists with ML and AI because of the high degree of nonlinearity and the collinearity of a model's features. While one predictor in combination with another predictor might not exacerbate discrimination, that same variable in combination with a different set of predictor(s) might result in a sudden jump in impact on a protected class. Thus, you cannot examine the impact of each variable, but must examine the impact of combinations of variables. If you have just 30 predictor variables, there are over 1 billion alternative specifications that might lead to a less discriminatory model. As a result, regulators are concerned that banks will not be able to devote the resources required to find a less discriminatory model. If this is true, then machine learning itself may be more discriminatory than GLMs.

Measuring Discrimination

When checking for disparate treatment, banks are well aware that they may not include variables that explicitly control for protected class status. Because of that, questions generally come down to possible proxies for protected class status, as well as the effect of geographical characteristics in lending models. To some extent there is a "we'll know it when we see it" approach to determining the legitimacy of variable selection. Fair lending compliance teams typically review variables to be included in a model one-by-one, flagging potentially troublesome variables for further analysis. If a variable seems suspicious, they usually look for correlations with protected class status, rejecting that variable if it appears too highly correlated. They also reject any variable that has no clear relationship with the outcome except for its disproportionate relationship with protected class status.

Disparate Impact testing typically requires more statistical analysis. Testing is done by calculating whether protected classes are, on average, disadvantaged by the model's output *vis-à-vis* comparator groups. This is a relatively simple process, usually using one of two metrics. When the model is to be used as a dichotomous "yes/no" offer selection, then the Adverse Impact Ratio ("AIR") is used. The AIR is known in statistics as the Relative Risk Ratio, and is defined as:

$$AIR_{pc} = \frac{\%\ Accepted_{p\,c}}{\%\ Accepted_{cc}}$$

When the output will be used as a continuous variable, then we suggest the use of the Standardized Mean Difference ("SMD"). More commonly known as "Cohen's d", this statistic is a general measure of effect size that can be compared across outcomes due to its scaling.

$$SMD_{pc} = \frac{100 * (\hat{x}_{pc} - \hat{x}_{cc})}{\sigma}$$

In these formulae, pc represents the protected class; cc is the comparator class (e.g., males in a female/male analysis); $\hat{x}$ is the average outcome for each of the groups; and σ is the standard deviation of the population.

The existence of discrimination requires not only statistical significance, but also a consideration of "practical significance." Practical significance is a threshold that is set by a court or regulator to determine whether the difference is *meaningful*. The idea is that nearly any difference can be statistically significant when one is using thousands or millions of observations, but that difference may not reflect an outcome that is appreciably different for the protected class. A complication that banks currently face is that regulators have not yet specified how they measure practical significance. Our practice is to suggest a threshold of 0.80 or 0.90 for the AIR, and -30.0 for the SMD.

If we do see a practically significant difference, we then ask, "Is the model a valid predictor of the outcome being forecast?" Further, if so, "Can we find an equally predictive model that has less of an impact?" In GLMs, this is a straightforward process: the banks run alternative model specifications and then test the disparate impact and quality of the models. If the models improve disparate impact while maintaining model quality, then the bank is required to adopt that model.

The Future of Machine Learning and Regulatory Compliance

Regulators' concerns regarding fair lending compliance grow from the "black box" nature of many ML algorithms. To the extent that we can better explain the predictions from these models, we will be able to allay some of their concerns. We will be able to further assuage their concerns by having a robust and effective alternative model building process when a model shows evidence of disparate impact.

In order to help regulators become more comfortable with machine learning, we believe the most important areas for work are: (1) bringing interpretability into machine learning decisions, and (2) providing banks with ways to test for less discriminatory models without major business interruption or high cost.

To address the problem of interpretability, we are implementing algorithms such as K-LIME in order to determine the extent to which each variable in the model has contributed to the model's decision. We also use other metrics, like variable importance, to give broader insights into what drives model decisions. In essence, we use information from these metrics to help regulators "get under the hood" of the models, so that they can be convinced that it is not discrimination that is driving decisions.

With regard to improving disparate impact through alternative model building, our experience has shown that the only feasible way to achieve these goals in machine learning is to apply AI to the AI models themselves. In other words, AI is used to fix AI. Here, we use a combination of AI algorithms, high powered computing, and smart search routines in combination to find the best model, both from a predictive quality and a fairness sense. We have found that these methods can be provided in a cost and time effective manner.

In sum, regulators' concerns regarding fair lending compliance have focused on the fact that it is very difficult to understand why most machine learning algorithms make a particular decision. We believe that banks that are able to explain their models and show that they have performed a robust search for the least discriminatory model will be able to pass regulatory challenges. Fortunately, both in industry and academia, the challenges of interpretability are being met, which means that banks that implement these methods will be able to overcome the regulatory hurdles that they currently face.