The Foerder Institute for Economic Research
at Tel Aviv University

The Eitan Berglas School of Economics

מכון למחקר כלכלי על שם ד"ר ישעיהו פורדר
על יד אוניברסיטת תל אביב

בית-הספר לכלכלה ע"ש איתן ברגלס

A Note on the Equilibrium Effects of Predictive Credit-Risk Models

Ran Spiegler

Discussion Paper No. 3-2023

The Foerder Institute for Economic Research
and
The Sackler Institute of Economic Studies

A Note on the Equilibrium Effects of Predictive Credit-Risk Models*

Ran Spiegler†

March 29, 2023

Abstract

I present a simple model of a credit market in which lenders use predictive models to evaluate borrowers' credit risk. Each firm trades off its ability to predict borrowers' risk according to their observed characteristics against their simplicity. Firms are heterogeneous in the weights they attach to each conisderation. Crucially, firms evaluate risk models' predictive success against the aggregate distribution of active borrowers. I show that in this model, lenders that attach low importance to explainability exert a positive externality on other lenders, because their complex predictive models make the aggregate distribution of active borrowers less adversley selective.

*Financial support from the Sapir Center and the Foerder Institute is gratefully acknowledged. I also thank Nir Rosenfeld and Dan Zeltzer for helpful discussions.

†Tel Aviv University and University College London. URL: https://www.ranspiegler.sites.tau.ac.il/. E-mail: rani@tauex.tau.ac.il.

1 Introduction

Recent years have seen the growing popularity of "machine learning" prediction algorithms. The driving force behind this development is the increasing availability of large datasets that contain numerous personal characteristics of consumers, job candidates, borrowers etc., as well as choices they make in various settings. The new technologies improve the ability to predict agents' future outcomes (consumption decisions, work performance, solvency) on the basis of their observed characteristics.

At the same time, critics bemoan the "black box" aspect of prediction algorithms — namely, the complexity and inscrutability of how they map observable characteristics to predictions. This criticism has led to calls for "explainable AI", which would incorporate simplicity as a criterion for evaluating predictive models. A different critique of prediction algorithms in contexts such as credit markets is that they condition on personal characteristics in a way that effectively leads to unwarranted discrimination (see O'Neil 2017), Kleinberg et al. (2018), Gillis and Spiess (2019), Gunning et al. (2019) and Kearns and Roth (2019) for discussions of these and other critiques of the growing use of predictive algorithms).

In this note, I explore some implications of the trade-off between prediction algorithms' predictive power and simplicity, in the context of a very simple model of a credit market. In the model, lenders try to predict potential borrowers' credit risk; the implicit price of a loan to a given borrower reflects this prediction. Lenders choose between two predictive models: A complex model that conditions on borrowers' characteristics and perfectly predicts whether they would default on a loan; and a simple model that does not discriminate between borrowers. The simple model relies on aggregate data regarding the solvency of borrowers. Crucially, however, the data is restricted to *active* borrowers — namely, borrowers who are granted loans. This data is potentially selective, yet the simple model does not correct for sample selection. Each lender is characterized by a parameter that determines how it weighs models' predictive power (evaluated by their mean squared error) and their complexity.

The key insight is that since all lenders evaluate predictive models against aggregate, selective data, they exert an externality on each other. In particular, a lender that places a large weight on model simplicity will tend to adopt the simple predictive model. Adoption of a simple model that fails to discriminate between borrower types generates an adverse selection effect, which means that the population of active borrowers becomes more selective. Since this selective population is more uniform, its credit risk is more predictable even with a simple model, which means that other lenders will be impelled to adopt the simple model. This externality can exacerbate the forces that lead to Akerlovian breakdown of the credit market. Put differently, simplicity of predictive models (whether interpreted as explainability or as a distaste for discrimination) has a novel externality that affects the performance of credit markets.

This note is related to a few strands in the literature. Fuster et al. (2020)) is an empirical study that evaluates the effect of fine predictive algorithms on the performance of credit markets, focusing on the issue of discrimination. Pacciano et al. (2021) is an example of a computer-science paper that acknowledges the problematic reliance of credit-risk predictive models on selective data, and offers an ad-hoc method for addressing it. Jehiel and Mohlin (2021) explore the role of endogenously coarse subjective models in selection markets, from a different angle. Finally, Eliaz and Spiegler (2019,2022) study the effect of simplicity-seeking prediction algorithms on the reporting incentives of agents who interact with such algorithms.

2 A Model

Consider a market that consists of a continuum of consumers (a.k.a borrowers) and firms (a.k.a lenders). Let X be a finite set of consumer types, and their distribution in the market is denoted p. The interpretation of a type is that it can be described by a collection of observable characteristics. Each type is associated with a number in $\{0,1\}$, which indicates the type's credit risk. For our purposes, we can identify consumer types with their credit risk, and hence we can assume that $X = \{0,1\}$. For any probability distribution

q over X, denote its expected value conventionally by E_q.

On the other side of the market, each firm is associated with a number $c \geq 0$, which captures how the firm trades off its ability to predict consumers' credit risk by their observable characteristics against the loss in simplicity or explainability that this entails. I will make this trade-off precise below. The distribution of values of c in the market is distributed according to some atomless *cdf* $F[0, \bar{c}]$.

I will now define a notion of market equilibrium, which aims to capture the following process. Firms try to predict consumers' credit risk, according to their predictive model and aggregate empirical data about consumers' loan-repayment performance. Firms use this prediction to price a loan to any individual consumer, potentially as a function of his observable characteristic. The implicit pricing has the property that consumers whose actual risk is below their predicted risk are excluded from the firms' pool of borrowers. This means that the data that is available to firms is selective: firms only have data about the performance of consumers who are granted loans. Furthermore, the data exhibits *adverse* selection, because of the exclusion of consumers whose credit risk is overestimated by firms' predictive models. Each firm evaluates the two possible predictive models by their ability to account for credit-risk variation in the data, as well as by their simplicity/explainability. Note that firms make use of *aggregate* data; we will revisit this assumption below.

Definition 1 (Market equilibrium) *A market equilibrium is a pair* (p^*, c^*)*, where* $p^* \in \Delta(X)$ *and* $c^* \geq 0$*, such that:*

$$\begin{aligned} p^*(x) &= F(c^*) \cdot p(x) + (1 - F(c^*)) \cdot p^*(x \mid x \geq E_{p^*}(x)) \\ c^* &= \sum_x p^*(x)[x - E_{p^*}]^2 \end{aligned}$$

The interpretation is as follows. The object p^* describes the selective distribution of the types of *active* consumers — namely, those who are granted loans in equilibrium. A predictive model that makes use of consumers' observable characteristics perfectly predicts their credit risk. In contrast, a

simpler predictive model that neglects this information can only make an average assessment, based on the selective data. When choosing between these two predictive models, firms trade off their predictive power (measured by their mean squared error) against their complexity. In equilibrium, firms opt for the complex model if and only if their complexity cost is below c^*. When a firm uses the complex model, the type distribution of their customer base matches the distribution in the general population, because their implicit pricing does not induce adverse selection. In contrast, when a firm uses the simple model, it effectively excludes consumers whose credit risk below is below the level the model predicts — namely, the average level in the selective aggregate population of active consumers.

The trade-off between predictive models' mean squared error and complexity is reminiscent of regularized regression methods such as Lasso (Tibshirani (1996)). However, note that the trade-off is based on the actual equilibrium distribution, rather than on a sample from it. Therefore, while regularization is an attempt to address the overfitting problem of model selection in the presence of a finite sample, the simplicity criterion that underlies the equilibrium definition captures a concern for explainability of predictive models, or a distaste for the discrimination they entail.

3 Analysis

This section analyzes market equilibria in the model. We saw in Section 2 that market equilibrium may exhibit some amount of adverse selection, if some firms adopt the simple model. As often is the case with market models in the presence of adverse selection, the present model has an equilibrium with total market breakdown — i.e., the only active consumer types are those with the highest credit risk $x = 1$. To see why this is an equilibrium, note that by definition, this class of consumers forms a homogenous population: they all have the same credit risk. Therefore, even the simple predictive model, which does not bother to take consumers' observable characteristics into account, will have zero mean squared error. As a result, all firms will opt for this model and predict a credit risk of $x = 1$. As a result, no other consumer

type will choose to be active, thus sustaining the equilibrium. Because of its conventional Akerlovian logic, I refer to this market equilibrium as the trivial equilibrium. The question is whether there are equilibria with broader consumer participation. The following result provides a sufficient condition for a negative answer.

Proposition 1 *Suppose* $F(c) \geq \frac{1}{2} - \sqrt{\frac{1}{4} - c}$ *for all* $c \in [0, \frac{1}{4}]$. *Then, the unique market equilibrium is the trivial equilibrium.*

Proof. For a non-trivial equilibrium to exist, there has to be some c^* with $F(c^*) > 0$, such that every firm with $c < c^*$ opts for the complex predictive model. By the definition of market equilibrium,c^* is equal to the variance of p^*. Denote $\alpha = F(c^*)$, and let β be the probability of $x = 1$ according to p. Since $p^*(x \mid x \geq E_{p^*}(x)) = 1$ for $x = 1$,

$$E_{p^*}(x) = p^*(x = 1) = 1 - \alpha(1 - \beta)$$

Therefore,

$$Var_{p^*} = \alpha(1 - \beta) \cdot [1 - \alpha(1 - \beta)]$$

By the definition of equilibrium, $c^* = Var_{p^*}$. It follows that

$$\alpha = F\{\alpha(1 - \beta) \cdot [1 - \alpha(1 - \beta)]\}$$

such that

$$\alpha(1 - \beta) < F\{\alpha(1 - \beta) \cdot [1 - \alpha(1 - \beta)]\}$$

Yet, the condition that $F(c) \geq \frac{1}{2} - \sqrt{\frac{1}{4} - c}$ for all $c \in [0, \frac{1}{4}]$ implies that $z \geq F(z(1 - z))$ for every $z \in [0, 1]$, a contradiction. ■

The significance of this result is that if the distribution of c is not too concentrated in the low range, total market breakdown is a necessary equilibrium outcome. The reason is a novel externality that the model captures. Observe that the two equations that define a market equilibrium can be combined into a single equation:

$$p^*(x) = F(Var(p^*)) \cdot p(x) + (1 - F(Var(p^*)) \cdot p^*(x \mid x \geq E_{p^*}(x)) \qquad (1)$$

This is due to the fact that the R.H.S of the equation that defines c^* is, by definition, the variance of p^*. The weight that the R.H.S of (1) puts on the adversely selective conditional distribution $p^*(x \mid x \geq E_{p^*}(x))$ decreases with this variance. Since the selective distribution itself has lower variance than the general, unconditional distribution, we have a positive-feedback effect that can result in total market breakdown.

The key intuition here is that as the aggregate empirical distribution p^* becomes more adversely selective, it also has lower variance, which implies a larger share of firms that adopt the simple predictive model (because the loss in predictive power from adopting this model goes down), which in turn strengthens the adverse-selection pressure. In other words, firms with low c exert a positive externality on firms with high c: their adoption of the complex model makes the aggregate distribution of active consumers less selective and therefore more varied, thus enhancing the force that impels other firms to adopt the complex predictive model, too.

A no-externality variant

To get a better understanding of the latter point, consider a variant on our model, in which each value of c defines a separate market. In this version of the model, firms do not use aggregate data to evaluate the simple predictive model. Instead, it focuses on the distribution of active consumers at firms with the same value of c.

Thus, the reduced equation (1) that describes equilibrium in the original model is transformed into the following equation:

$$p_c^*(x) = \mathbf{1}(Var(p_c^*) > c) \cdot p(x) + \mathbf{1}(Var(p_c^*) \leq c) \cdot p_c^*(x \mid x \geq E_{p_c^*}(x))$$

where p_c^* is the distribution of active consumer types at firms with a given c. We can see that only when $Var(p) > c$, there is a solution to this equation for which $p_c^* = p$ — i.e., the firm adopts the complex model, hence its population of active consumers coincides with p, which is self-sustaining only if the variance of p exceeds c.

It follows that in this version of the model, the equilibrium fraction of firms that adopt the complex model can be as high as $F(Var(p))$. This

means that some consumers with low credit risk are active in equilibrium. By comparison, when F satisfies the condition in Proposition 1, no firm adopts the complex model in equilibrium, which means a unique, trivial equilibrium.

4 Conclusion

This note presented a simple example of a credit market, in which firms use predictive models to evaluate borrowers' credit risk. Different firms use different weights to trade off the model's predictive ability against its simplicity (interpreted as explainability). The key assumption in the example was that firms evaluate this trade off against aggregate data on active borrowers. This means that the data is also selective — i.e., it is restricted to borrowers who are granted a loan. As more firms adopt a complex model that enables finely tailored credit pricing, the aggregate data becomes less adversely selective, which also means that it has lower variance and thus makes complex models even more valuable for other firms. This positive externality generated by complex predictive models is this note's main insight. Of course, this insight is valid only to the extent that the relative advantage of complex models increases when the market becomes less adversely selective. Examining the relevance of this connection is left for future research.

References

[1] Eliaz, K. and R. Spiegler (2019), The Model Selection Curse, American Economic Review: Insights 1, 127-140.

[2] Eliaz, K. and R. Spiegler (2022), On Incentive-Compatible Estimators, Games and Economic Behavior 132, 204-220.

[3] Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai and A. Walther (2022), Predictably Unequal? The Effects of Machine Learning on Credit Markets, Journal of Finance 77, 5-47.

[4] Gillis, T and J. Spiess (2019), Big Data and Discrimination, University of Chicago Law Review 86, 459-488.

[5] Gunning, D., M. Stefik, J. Choi, T. Miller, S. Stumpf and G. Yang (2019), XAI—Explainable artificial intelligence, Science Robotics 4, essay 7120.

[6] Jehiel, P. and E. Mohlin (2021), Cycling and Categorical Learning in Decentralized Adverse Selection Economies, mimeo.

[7] Kearns, M., & Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press.

[8] Kleinberg, J., J. Ludwig, S. Mullainathan and C. Sunstein (2018), Discrimination in the Age of Algorithms, Journal of Legal Analysis 10, 113-174.

[9] O'Neil, C. (2017), Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Broadway Books.

[10] Pacchiano, A., S. Singh, E. Chou, A. Berg and J. Foerster (2021), Neural Pseudo-Label Optimism for the Bank Loan Problem, Advances in Neural Information Processing Systems 34, 6580-6593.

[11] Tibshirani, R. (1996), Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B (Methodological), 267-288.