Algorithmic Credit Scoring and FICO's Role in Developing Accurate, Unbiased, and Fair Credit Scoring Models

Dr. Bernard Siskin[1]
BLDS, LLC and Chairman of the Board at SolasAI
Washington, DC, USA
bsiskin@bldsllc.com

Nicholas Schmidt
CEO of SolasAI and the AI Practice Leader at BLDS, LLC
Philadelphia, PA, USA
nicholas.schmidt@solas.ai

Dr. Bryce Stephens
Principal at Bates White Economic Consulting
Washington, DC, USA
bryce.stephens@bateswhite.com

1. INTRODUCTION

The way in which lenders make decisions about whether to extend credit to consumers has changed drastically over the last few decades. Lenders no longer use face-to-face meetings to subjectively assess a consumer applicant's likelihood of default. Instead, lenders employ analytical methods that enable them to objectively determine the risk of lending to a consumer. In turn, the consumer's ability to understand why he or she has been approved or denied a loan has increased too because lenders are now able to more thoroughly explain why a given loan was granted or denied.

These objective analytical methods were pioneered by Fair Isaac Corporation ("FICO"). Founded in 1956, FICO developed credit scoring models that consider only credit history and enable lenders to more accurately predict whether a consumer will default on a loan, which helps eliminate subjective bias and expand credit access to more consumers. These more accurate models enable lenders to underwrite a larger number of borrowers while controlling for risk; in other words, more accurate models drive expanded access to credit.

FICO became a household name with the advent of the "FICO Score," first released in 1989. In 1995, with the FICO Score already in widespread use in the consumer lending industry, Fannie Mae and Freddie Mac – the government sponsored enterprises that guarantee most residential mortgage loans in the United States – began using FICO Scores to assess the credit quality of the loans they purchased. FICO Scores have continued to be implemented and improved, and they are now used across the financial industry as nationally accepted objective measures of the default risk of consumer borrowers. Because of their wide-reaching use and impact, FICO Scores are monitored not only by lenders, investors, and regulators, but also by consumers themselves.

In the last several years, however, the rise of machine learning models, as well as the use of alternative data in those models, have led to questions about whether algorithmic credit scoring can be fair to all consumers[2]. The focus has shifted to issues of fair lending accountability and the need to evaluate whether credit scoring models may cause or perpetuate bias against protected classes – especially racial minorities.[3] These concerns are being raised with respect to credit scores like the FICO Score that are built on traditional credit bureau data using longstanding and accepted statistical techniques, as well as new credit scoring models built on alternative data using artificial intelligence and machine learning techniques. Congress, regulators, consumer groups, and financial institutions have increased their focus on social justice and fair lending practices, while proponents of machine learning models that consider alternative data claim improvements over existing credit score modeling techniques.[4]

This paper examines these questions as they relate to the FICO Score. Section 2 provides background on the development of algorithmic credit scoring and the FICO Score. Section 3 discusses the fair lending concerns presented by credit scoring models such as the potential for proxy bias, where model variables identify protected demographic groups, and for prediction bias, where a model treats such groups less favorably than the population as a whole. The remainder of Section 3 then reviews how FICO has addressed and continues to address fair lending with the FICO Score. Section 4 examines the recent growth in machine learning models and their application to credit scoring, particularly those models that consider alternative data, and the fair lending issues that arise in comparison to the FICO Score. Finally, Section 5 provides a summary of our findings with respect to the FICO Score.

We conclude that the FICO Score considers only neutral, objective factors predictive of credit performance based on a consumer's credit bureau file, and it does not consider any variables that are proxies for protected groups. We also conclude that the FICO Score contains no evidence of prediction bias that causes studied protected groups to score lower than the overall population, after controlling for true default rates. Finally, with respect to machine learning models, we believe that recent advances have enormous potential for predicting human behavior in beneficial ways. However, in the context of credit scoring, current machine learning techniques, particularly those that consider untested or insufficiently studied alternative data, raise transparency and fair lending concerns that are not present in the FICO Score.[5]

2. FICO AND THE DEVELOPMENT OF CREDIT SCORING MODELS

Today, credit scoring models are used by lenders across the financial services industry to assess the likely financial performance of consumer borrowers. While FICO Scores are the most well-known, nearly every lender uses data and additional algorithms – often in conjunction with FICO Scores – to gain further insight into their consumer customers and prospects to identify the extent to which a consumer represents a credit risk to an institution or credit product.

Different FICO Scores have been developed in the 32 years since the original score was released. The original FICO Score and its subsequent versions consider only a consumer's credit information maintained at the three nationwide consumer reporting agencies (CRAs), Equifax, Experian, and TransUnion. By providing lenders with a tool for evaluating consumer credit risk objectively, based on neutral credit risk factors, the original FICO Score reduced the influence of subjective human judgment or biases (with respect to, for example, race, gender, or family status) that were often present in manual underwriting, which helped to expand access to credit to more consumers. Subsequent versions of the FICO Score have further expanded consumer credit access by increasing model accuracy through the extraction of new insights from traditional CRA data and through the consideration of the evolution in data trends and changes in consumer loan products. While FICO has developed new credit scores like FICO Score XD and UltraFICO Score that also consider alternative data to expand the scorable population safely and reliably, this paper focuses on the "flagship" FICO Score that considers only traditional CRA data.[6]

There are several key aspects of the FICO Score that are important to this paper. First, the FICO Score is based solely on credit information available in the consumer's credit file at the CRAs. As a result, the FICO Score does not consider macroeconomic factors or other factors that may be outside the control of the consumer that may impact default at any point in time and may change over time as economic conditions change. Second, the FICO Score is designed to be transparent and palatable[7], because of applicable legal requirements[8] and FICO's commitment to educating consumers about credit scores[9]. Third, the objective of FICO's modeling process is to accurately rank order consumer risk of default across many types of loans. In our experience, most algorithms are built using data from at least tens of thousands of people; the FICO Score, on the other hand, is built using credit histories of millions of consumers that come from multiple institutions, on a representative national sample basis, allowing a lender to obtain a more comprehensive assessment of the default risk of potential new customers.

Rank ordering means that, for a given lending product, a consumer with a higher score will be less likely to default relative to a consumer with a lower score, which is valuable information to a lender[10]. The FICO Score is designed to maintain its rank ordering properties across time and macroeconomic environments – even if the underlying probability of default changes for a given FICO Score. For example, during an economic downturn, the probability of default for consumers with given credit scores may increase relative to the probability of default during better economic conditions[11]. However, the rank ordering is consistent: someone with a FICO Score of 650 generally remains a more likely default risk than someone with a score of 700.

Since its creation, the FICO Score has brought several advantages to consumers and to the financial services industry. First, it expanded access to credit to more consumers by reducing the risk that the subjective biases, beliefs, and misconceptions of an individual loan officer or underwriter could adversely affect a lending decision. It also provided significant business advantages to the industry: credit score models can – and do – utilize far more information in determining default risk than any individual human being could consider. The result is a credit score that is both objective and more accurate (i.e., predictive of default).

However, while credit score algorithms help reduce the influence of subjective bias, such models may introduce the possibility of algorithmic bias against protected classes if the models are not designed carefully and thoroughly tested for bias. This concern is addressed in the sections that follow.

3. FAIR LENDING CONCERNS FOR CREDIT SCORING MODELS

A. Two Important Fair Lending Concerns: Proxy Bias and Prediction Bias[12]

One major legal concern with credit scoring models is that the model may use protected class status as a criterion by using a variable or combination of variables that act as a proxy for protected class status. That is, the predictive power of the variable or variables may be primarily due to the ability to proxy protected class status resulting in what is referred to as proxy bias.

A second major concern is whether the model causes a protected class to score lower than the overall population due to prediction bias against protected class members. Prediction bias presents a fair lending concern with respect to such a protected class when the model scores that protected class lower than the population as a whole even after controlling for true default rates. In other words, the model assigns the same score for the protected class members and the members of the population as a whole, but the true likelihood of default of the protected class is actually lower than what the model predicts. In such a case, the model's output would be biased against the members of that protected class.

To illustrate, consider the following example of prediction bias. Suppose we look at consumers in a protected class for which the score distributions of the model skew lower on average than the overall population. When comparing the actual credit performance of members of the protected class to the members of the overall population who have the same predicted likelihood of default (i.e., who score the same), the likelihood of default should be the same on average. However, if the protected class borrowers actually default at a significantly lower rate than similarly scoring borrowers in the overall population, then this would represent a biased model adverse to that protected class. Such prediction bias would harm the protected class borrowers relative to the overall population because loan rejection rates or interest rates for protected class members based on this model would likely be higher than they would have been absent such prediction bias.

B. Fico Scores have been Shown Not to Present Proxy Bias Concerns

As explained above, the FICO Score only considers CRA data and only uses variables that the individual may have the ability to affect based on the individual's financial behaviors over time. By restricting the data to only CRA data, the FICO Score only considers variables that have sufficient coverage in the total consumer population[13]. Further, FICO's model development process includes the involvement of experienced data scientists – a "human-in-the loop" – to ensure that each variable is reviewed for a direct and logical connection to risk of default.[14]

FICO follows a rigorous model development and validation process. Model input variables are carefully selected to ensure that they are neutral, objective, and predictive of credit performance. The FICO Score does not consider personal identifying information, demographic data, geographic location, or any legally prohibited bases or factors under fair lending laws. It constitutes an "empirically derived, demonstrably and statistically sound" (EDDSS) credit scoring system under Regulation B which implements the Equal Credit Opportunity Act ("ECOA").

In addition, the FICO Score does not consider any factors that are strong proxies for prohibited bases under fair lending laws or that generate their predictive power primarily by separating the protected class members. This is supported by an independent, empirically supported research study by the Federal Reserve Board.[15] Those researchers studied the potential for proxying protected class status for each of the 312 CRA "credit variables." Using a methodology that attempted to emulate the process for selecting variables from CRA data that are predictive of consumer credit default risk, the researchers isolated and tested the impact on the credit scores caused by the correlation of the 312 variables to protected class status compared to its impact independent of protected class status.[16] They found no evidence that any of the 312 credit variables were meaningful proxies for race, ethnicity, gender, or age.[17]

FICO's own testing of the FICO Score is consistent with the Federal Reserve Board study and confirms that the FICO Score does not consider any variables that are proxies for any protected groups. These testing results are consistent with our understanding of credit scoring models that likewise carefully select neutral variables and consider only CRA data, based on our lengthy experience in working with lenders that use such credit scores and in connection with the management and validation of such models.[18]

C. FICO Scores Have Been Shown Not to Present Prediction Bias Concerns

The second fair lending concern is whether the model scores a protected class lower than the overall population because of prediction bias in the model against the protected class. This prediction bias can be caused by bias in the accuracy and completeness of the development data or by the possibility that variables included in or omitted from the model may impact predictions differentially based on protected class status. Even the use of highly accurate and complete data may lead to model prediction bias and raise fair lending concerns.

FICO does not consider any prohibited bases (e.g., race, ethnicity, gender, or age) during development of the FICO Score. Importantly, the development data used by FICO modelers is depersonalized such that it does not contain any personally identifying information. As discussed above, concerns about bias because of any inaccuracy or incompleteness of the development data are reduced by FICO's use of only CRA data. Under the Fair Credit Reporting Act ("FCRA"), lenders must ensure that consumer credit information furnished to the CRAs is accurate and complete, the CRAs are required to maintain and report accurate and complete credit information, and consumers have the right to dispute the accuracy of the credit information furnished by lenders and maintained by the CRAs. Further, FICO has taken steps to help educate consumers of their rights, providing information on how consumers can obtain their credit reports and how they can check the accuracy of their credit report data.[19]

Two causes of prediction bias against protected classes are often posited. One is caused by variables included in the model, and one is caused by factors excluded from the model. In the first case, the variables included in the model may impact members of the protected and other classes differently. Because a statistical model generally averages the prediction of default across the population being scored, controlling for all the other factors in the model, differences in the way a variable interacts with the likelihood of default across populations can lead to over- or under-prediction. For example, as a hypothetical, suppose that the likelihood of default increases for non-Hispanic whites by 5% for each additional open credit account (e.g., credit card, mortgage, auto loan, etc.) above some threshold. Now suppose that the likelihood of default only increases by 3% for African Americans for each additional open line. In that case, when race and ethnicity are excluded, the model will average out these effects and assign a higher likelihood of default for African Americans, and a correspondingly lower likelihood for non-Hispanic whites. Because non-Hispanic whites are the majority of the population, the impact would be greater on African Americans as a group than on non-Hispanic whites as a group.

The second posited cause of prediction bias is that there may be factors that are not included in the model that may impact the outcome differently by class. When such an omitted variable exists, the average marginal impact of that omitted variable is added to the level of default being predicted. That is, the effect of the omitted variable is assumed to be equally likely to impact all persons' likelihood of default. Hence it should not impact the ranking of default likelihood of the individuals. However, if the average value of the omitted variable were to differ by protected class status, then the average impact would be too low for a class with the better average value on the omitted variable and too high for a class with the worse average value on the omitted variable. This would result in overprediction of default for the classes with the better value on the omitted variable, and underprediction of default for the classes with the worse value on the omitted variable.

Clearly, there are factors that could influence credit outcomes that are omitted from FICO Score models (and all other credit scoring models) because some factors are unmeasurable or unavailable.[20] In our experience, however, those omitted factors that are positively correlated with one's risk of defaulting (such as the probability of losing one's job) are often positively correlated with protected classes for which credit score distributions frequently skew lower on average than the overall population, and those omitted factors that are negatively correlated with one's risk of defaulting (such as wealth) are often negatively correlated with those same protected classes. Therefore, we would not expect omitted variable bias to cause the model to overpredict the default rate of those protected classes.

FICO has tested the FICO Score overall for fair lending compliance. This testing includes measuring whether the FICO Score has a prediction bias adverse to the two protected classes for which the credit score distributions frequently skew lower on average than the overall population – African Americans and Hispanics. The analysis indicates that the FICO Score showed no evidence of prediction bias against those protected classes. In other words, the testing confirmed that – comparing persons with the same likelihood of repayment/default – the model did not score individuals in these protected groups lower than individuals in the population as a whole.

Again, this is consistent with our experience working with credit scoring models and working with lenders in connection with model risk management and validation. It is our view that FICO, as the developer of a generally applicable consumer credit score, is properly focused on building – and has built – a score that is based on neutral objective factors that are predictive of consumer credit performance, does not contain any variables that proxy for protected groups, and is not biased against protected groups whose score distribution skews lower than the overall population.

4. FAIR LENDING IMPLICATIONS OF NEW MACHINE LEARNING MODELS AND THE USE OF ALTERNATIVE DATA

In recent years there has been a rapid growth in the use of machine learning algorithms, as well as the use of alternative data in such models.[21] In particular, the early adoption and apparent success of many smaller FinTechs led many more traditional banks to consider using these technologies along with alternative data. In the last several years, larger and more established firms have explored putting these types of models that consider alternative data into production.

A. The Potential Black-Box Effect of Machine Learning Models and the Fair Lending Implications for Credit Scoring

Many types of machine learning algorithms can more easily be trained on data sets containing hundreds and even thousands of variables – which may include alternative data that is less complete, less studied, or less frequently used than traditional CRA data. Further, the flexibility of these models allows them to create intricate combinations or interactions of multiple variables. In fact, this is one of the primary appeals of machine learning algorithms: they can quickly and often effectively find patterns in data that are hidden or would take humans an extraordinary amount of effort to discover. At least in theory, this allows machine learning models to derive the most accurate predictive outcomes possible for highly complex datasets.

This flexibility, however, often comes at a cost. Highly flexible machine learning algorithms often have limited transparency: understanding a variable's contribution to a prediction, the way the variables interact with each other, and why the algorithm may have deemed the variable important may often be extremely difficult. In particular, while larger and more complex algorithms may generate better (i.e., more accurate) predictions, the inner workings of the models can become increasingly obscure. When these algorithms are particularly complex, they are often referred to as a "black box," meaning that while we understand what went into the algorithm (the data), and what came out of it (the predictions), the process by which the data are turned into predictions is obscured from view.

When considering transparency in machine learning, the ability for a person to understand a model is often broken into two separate concepts: explainability and interpretability. Interpretability is often defined as the ability to describe the inner workings of the model, whereas explainability is used to describe why a model gave a particular prediction.[22] Interpretability is especially important for model builders, businesses that rely on models, and, likely, regulators. These stakeholders want to understand whether a model is robust, reliable, and fair. Gaining an understanding of how a model handles data, weighs different factors, and creates predictions – the core of interpretability – allows these groups to gain comfort in the model. On the other hand, users and those affected by the model are most likely more interested in explainability. Here, someone rejected for a loan as a result of a credit scoring model might want to know – and likely has the legal right to know – why they were rejected and what they need to do to get accepted the next time they apply. This is where the accuracy of an explainability method is critical, often making explainable models preferable.

Significant progress has been made in recent years in creating methodologies that do an effective job interpreting and explaining many black-box machine learning algorithms.[23] However, despite this progress, some experts have argued that models for high stakes decisions should be limited to traditional or more interpretable machine learning models.[24] The goal of interpretable machine learning is to achieve the best possible balance between the flexibility of a machine learning algorithm with the requirements and desirable properties of an interpretable model which are typically straightforward to explain to those affected by models.

An example of an interpretable machine learning model is a class of models known as interaction generalized additive models, or GAMs. This interpretable model architecture might be considered near the edge of what is described as a machine learning algorithm, but, relative to traditional algorithms, they do achieve some of the additional flexibility of machine learning. In fact, these models represent a relatively small departure from FICO's own methodology used to build the traditional FICO Score.

FICO itself has developed interpretable machine learning models, including a neural network technique called "interpretable latent feature neural networks." This algorithm limits the complexity of the interactions of the inputs, which makes interpretability simpler. Further work from FICO has demonstrated how these models can be constrained to drive proper monotonicity requirements.[25] Despite this and other innovative work being done to break apart black-box algorithms to be interpretable, there are numerous other machine learning algorithms that many argue do not meet sufficient explainability standards. While true generally, the danger of using a model that cannot be sufficiently explained in the context of credit scoring may be especially problematic given fair lending laws and the federal regulations that require a lender to provide certain reasons why a customer was rejected for a loan or otherwise experienced certain negative changes in credit terms (i.e., an "adverse action notice").

FICO does not currently use its neural network model or other unconstrained machine learning algorithms in the calculation of the FICO Score. It has analyzed unconstrained machine learning techniques and does not yet see any significant improvement in predictiveness (i.e., accuracy) from using these techniques, and the cost in palatability and transparency currently outweighs the value of any marginal increase in predictiveness.[26]

However, FICO does use interpretable machine learning models during FICO Score development to identify powerful interactions and variables that are then captured in a refined FICO Score algorithm to increase predictiveness of the score. Further, FICO supports using interpretable machine learning algorithms – possibly in future credit scoring models – where it is shown to be safe and responsible to do so. In fact, outside of the credit scoring context, FICO already uses its sparse neural network model in its suite of models detecting fraud and money laundering.[27] FICO also supports the use of tested and reliable alternative data and uses machine learning techniques to gain insights into those alternative data sources (and variables). In the credit scoring context, FICO leveraged these machine learning techniques in the development of FICO Score XD and UltraFICO Score, which consider safe and reliable alternative data sources in addition to traditional CRA data to help expand credit access to consumers who cannot be scored using traditional CRA data alone.[28]

In the credit scoring context, then, a significant fair lending concern today relates to whether new black-box machine learning models – especially those utilizing large amounts of untested alternative data – may create interactions of variables that are proxying protected class status. The use of such proxies for protected classes could violate fair lending requirements under ECOA and the Fair Housing Act (FHA), which prohibit consideration of protected class status, or any factors that are strong proxies for protected class status, in the making of credit decisions.

As noted above, a proxy could include a variable or combination of variables whose predictive ability is predominately based on the ability to predict protected class status. Under this definition, the use of such a proxy would likely violate fair lending laws. Given that protected class status can be predictive of a consumer's risk of default on a loan, a machine learning model could attempt to use the available variables to proxy race if that proxy will increase predictiveness and accuracy.[29] The use of such a proxy could run afoul of applicable fair lending laws and perpetuate any societal discrimination that exists against the protected class. Such proxies may not be obvious in black-box machine learning models. Without significant and computationally complex efforts to search for such proxies, it is possible that the proxy could be hidden in the black box. Hence, the danger of unintentionally proxying protected class status in non-interpretable machine learning models has heightened fair lending concerns among regulators and consumer advocacy groups.

In contrast, as explained above, FICO places a very high importance on the transparency, explainability, and palatability of the FICO Score, consistent with applicable laws and with FICO's commitment to educating consumers about credit scores. The FICO Score is developed using well-accepted and time-tested methodologies designed to ensure the model meets these standards, and that the factors used are logically and directly related to consumer default risk.[30]

FICO's development methodology includes a "human-in-the-loop" design in which each data input is checked to make sure there is a solid justification for considering such a factor based on economic or financial theory. FICO's development methodology also captures complex interactions between variables using well-defined segmentation that is understandable and has been shown not to contain proxy or prediction bias based on protected class status. This is consistent with our experience in evaluating credit scoring models that consider CRA data, has been confirmed by FICO's own testing, and finds support in an independent, empirical study by researchers at the Federal Reserve.

B. The Accuracy of Machine Learning Models and the Fair Lending Implications

With the rise of machine learning, many question whether such models can be more accurate (i.e., predictive of default) than the FICO Score. A growing body of literature, however, has shown that any additional predictive power provided by complex and non-transparent algorithms may be overstated when compared to the advantages of their less complex and more transparent counterparts for consumers who can be scored with traditional CRA data.[31] This outcome is less surprising when one considers the context of credit scoring models: while machine learning algorithms are particularly good at finding hidden relationships within datasets where the data have not been extensively analyzed, CRA data has been analyzed extensively for decades by numerous experts. It is therefore likely that most meaningful relationships in the CRA data have already been discovered; moreover, those relationships appear fairly consistent over time.[32]

As noted, FICO has tested the extent to which using a machine learning methodology restricted to CRA data would increase the predictive accuracy of the FICO Score. Based on this testing, FICO found that the use of machine learning methodology would result in relatively little gain in predictive accuracy for this specific use case.[33] Further, beyond the question of accuracy, there are additional factors that need to be considered when determining if machine learning should be implemented. Caution must be taken because machine learning algorithms can be more susceptible to overfitting a model to the specific data seen during training, which means that when a model is put into production, it does not perform as well as expected. Additionally, such models often do not perform as well when used to score consumers who do not closely resemble the consumers on which the model was trained. Given these potential deficits – along with the possible loss in transparency, explainability, and palatability with some machine learning models, and the marginal increase in predictive accuracy – there is significant merit, at least in the context of credit scoring, to using the standard modeling techniques that FICO uses with the FICO Score.

It is important to note, however, that FICO does not ignore machine learning's ability to find predictive combinations of variables. FICO uses machine learning models such as Transparent Generalized Additive Model Tree (TGAMT), a model family which promotes explainability and transparency through design, to help identify potential predictive combinations of credit variables.[34] FICO then uses its "human-in-the-loop" methodology to see if the suggested combinations can be added to the scorecard model and still meet transparency, explainability, and palatability requirements. Based on our experience, FICO's use of machine learning to identify predictive combinations of variables during model design is an appropriate approach in the credit scoring context. This "human-in-the-loop" approach captures some of the benefits of machine learning without the fair lending and lack of transparency risks presented by some newer machine learning models for credit scoring.

5. CONCLUSION

In summary, the FICO Score has proven to be highly predictive of consumer credit default risk, without presenting the fair lending concerns that may be present in newer machine learning-driven models, particularly those that consider potentially unreliable alternative data sources. The FICO Score considers only neutral, objective factors available from traditional CRA data. The FICO Score has been shown not to consider any variables that are proxies for protected groups and not to contain any prediction bias that causes protected groups that have been historically disfavored to score lower than the overall population, after controlling for true default rates. Since the release of the original version in 1989, the FICO Score has helped to reduce subjective bias from consumer credit decisions and has led to more objective and accurate credit decisions which has increased access to credit for more consumers.

References

1. About the Authors

Dr. Bernard Siskin is the founder of BLDS, LLC and Chairman of the Board at SolasAI. He has specialized in the application of statistics in law, particularly in analyzing data for statistical evidence of discrimination. He has testified for both plaintiffs and defendants in more than 200 cases. Dr. Siskin has served as a statistical consultant to the U.S. Department of Justice, the Equal Employment Opportunity Commission, the U.S. Department of Labor, the Consumer Financial Protection Bureau, as well as numerous other federal, state and city agencies and Fortune 500 companies.

Nicholas Schmidt is the CEO of SolasAI and the AI Practice Leader at BLDS, LLC. Nicholas has spent over twenty years working at the intersection of law, regulation, and statistics, with more recent work focusing on how discrimination and bias can be identified and mitigated when companies employ algorithmic decisioning.

Dr. Bryce Stephens is a Principal at Bates White Economic Consulting. Prior to joining Bates White, Dr. Stephens was a Director at BLDS, LLC. He provides economic research, econometric analysis, and compliance advisory services, with a specific focus on issues related to consumer financial protection, such as the Equal Credit Opportunity Act (ECOA), and emerging analytical methods.

The authors thank FICO for providing access and support during the preparation of this white paper.

2. "Alternative data" is not uniformly defined. It is typically defined not by its scale or complexity, but by its novelty. For purposes of this paper, we define such alternative data to include: (1) data that is not normally used in credit models because it is not considered to be a direct measure of creditworthiness (though it may be a predictor of creditworthiness because of its correlation to factors that are direct measures of creditworthiness) and (2) data that may be considered to be a direct measure of creditworthiness but that is not reliably available. For example, data that would fit this definition includes social media interactions or educational attainment, which would fall into the first category, and income or cash flow, which would fall into the second category.

3. The Equal Credit Opportunity Act ("ECOA") (applicable to extensions of credit) and the Fair Housing Act ("FHA") (applicable to housing practices) prohibit lenders from considering prohibited bases such as race, gender, age, or familial status. As an exception, special purpose lending programs under ECOA may consider protected class or proxies for protected class. See https://bit.ly/3CkHA39.

4. Among other things, proponents of using machine learning and alternative data in credit scoring argue that their use adds coverage by scoring additional consumers who cannot be scored using credit bureau data alone. We agree that expanding coverage is a laudable goal, but not at the expense of transparency or fair lending. We anticipate releasing another paper on FICO's commitment to financial inclusion and further expanding coverage through alternative data that is safe and reliable.

5. As background for this paper, we were granted access to and conducted extensive interviews of a number of senior data scientists and model developers at FICO. Our conclusions in this paper are informed by our consideration of the information provided by FICO as well as our own extensive experience over several decades evaluating credit scoring models for fair lending compliance.

6. FICO Score XD considers a consumer's bill payment data and select public records data (https://www.fico.com/en/products/fico-score-xd) and UltraFICO Score considers a consumer's checking and savings account data (https://www.fico.com/en/products/ultrafico-score).

7. Palatability generally refers to whether the results of a model are acceptable or reasonable – i.e., the score should increase or decrease in a way that makes intuitive sense – based on a consumer's credit history and behavior. Palatability is important to consumers, lenders, and regulators.

8. ECOA and the Fair Credit Reporting Act ("FCRA") require lenders to provide certain disclosures to consumers when making adverse credit decisions. FCRA also provides consumers with the right to challenge and correct the credit information furnished by lenders and maintained by the CRAs. FICO provides outputs from the FICO Score model, known as reason codes, designed to help lenders meet their adverse action notice requirements under ECOA and FCRA by providing transparency to the credit underwriting process and explanation of the reasons adverse action was taken by the lender.

9. FICO makes a significant investment and effort in educating consumers on what a FICO Score means, what it considers, and how consumers can manage their financial health. FICO's "Score A Better Future" program (https://www.fico.com/en/sabf) is designed to bring together consumer advocates, credit educators, and community leaders at free educational events to help consumers – particularly lower income communities, first time home buyers, millennials, and those with no credit scores or less strong credit scores – to improve their understanding of FICO Scores and their overall financial health. In addition, myFICO.com provides tools, including certain free resources, to help consumers obtain and understand their credit report and FICO Scores (https://www.myfico.com/credit-education). Finally, the "FICO Score Open Access" programs allow lenders (https://www.fico.com/en/products/fico-score-open-access) and credit and financial counselors (https://www.fico.com/en/newsroom/fico-makes-fico-scores-available-financially-struggling-consumers-through-non-profit) to share FICO Scores directly with consumers with no fees from FICO.

10. The FICO Score is designed to maximize the accuracy of rank ordering of relative consumer risk. The actual level of risk of default for a given FICO Score value will vary by product as some products are inherently riskier than others and will also vary due to changes in macroeconomic conditions that can impact risk of default differently by product.

11. FICO also has developed a score known as the FICO Resilience Index ("FRI"), which is designed to be used in conjunction with FICO Score to capture how downturns in the economy are likely to affect the default risk of individual consumers. See https://www.fico.com/blogs/fico-resilience-index-now-available-lenders-pilot.

12. Here we focus on two types of bias that are often discussed when considering whether there may be discrimination resulting from the use of algorithms for credit scoring. However, there are other definitions of bias and discrimination that are often raised in the literature. Many of these focus on whether the use of variables is discriminatory even if the variables are predictive of default and not directly tied to protected class status. Further concerns are raised regarding the fairness of using any variables or algorithms that may result in any differences across protected classes. These definitions and the arguments about them are outside of the scope of this paper.

13. Entities that furnish data about consumers to CRAs have a legal obligation under the FCRA to furnish information that is accurate and complete, and to investigate consumer disputes about the accuracy of the furnished information and correct inaccurate information.

14. For more detail see the FICO white paper available at https://www.fico.com/en/latest-thinking/white-paper/introduction-model-builder-scorecard.

15. "See Does Credit Scoring Produce a Disparate Impact?" Robert B. Avery, Kenneth P. Brevoort and Glen Canner, Federal Reserve System, October 12, 2010, available at https://bit.ly/30pB79U.

16. Although the Federal Reserve Board researchers obviously could not specifically replicate FICO's methodology, or any other commercially available credit methodology, they do show that their model is highly correlated with two commercial credit scores that were available for their sample of data.

17. According to the authors, a variable is a proxy for protected class if its predictive power comes primarily from its correlation with the protected class and not from its correlation with default controlling for protected class status.

18. For example, banks that are subject to OCC oversight are required to periodically validate the models they use in credit decisioning. See OCC 2011-12 Supervisory Guidance on Model Risk Management, available at https://www.occ.treas.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf.

19. See endnote 9.

20. In contrast to any omitted variables which would not be considered by the model at all, there are variables that are considered by the FICO Score, but where the data is incomplete or unavailable at the CRAs, such as rental data.

21. The concepts and use of the terms "alternative data" and "machine learning" are often conflated and confused, where it is implied that machine learning requires alternative data or alternative data requires machine learning. While this may be due to their concomitant rise in popularity and usage, they are two entirely separate things. In this context, machine learning generally refers to classes of algorithms used to process data and create predictions, whereas alternative data refers to data sources and variables that have not traditionally been utilized for credit decisioning or are less obviously related to credit outcomes. Machine learning algorithms can use alternative data, but they do not require alternative data. Similarly, traditional algorithms can be used to process alternative data. However, machine learning algorithms are often used when creating credit models using alternative data for a number of reasons, including that many machine learning algorithms tend to find patterns in less studied data faster than could be achieved by a human.

22. Hall, et al (2021). A United States Fair Lending Perspective on Machine Learning. Frontiers in Artificial Intelligence. https://bit.ly/3wTNSpx. In addition, there has been significant research over the last few years on the development of quantifiably fair machine learning algorithms, whereby desired notions of fairness are improved, or satisfied, "by design" of a preprocessing, a training, or a post-processing algorithm. FICO is actively exploring innovations in this area, including working on extensions of its "TGAMT" technology discussed at page 12 and in endnote 34.

23. These attempt to explain individual predictions which can be used to understand models and, potentially, to form the basis of the adverse action notices required under ECOA. Likely the most used methods in credit scoring are based on the SHAP technique, which itself is derived from an approach created in the field of economic game theory. SHAP Values attempt to decompose a given prediction into constituent parts, showing how much each model feature drove the prediction up or down relative to some reference point. Locally Interpretable Model-Agnostic Explanations, or LIME, is another technique commonly used. While it is typically less precise than SHAP it has other beneficial properties that make it easy to use and interpret. Finally, another popular approach is Integrated Gradients, which can provide explanations for especially opaque deep neural networks.

24. For example, see Rudin, C. (2019), "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, 1(5), 206-215.

25. Monotonicity constraints are limitations that the user puts on the algorithm, forcing the algorithm only to consider a variable if it causes the prediction to go in a certain direction. For example, one might force debt-to-income to have a positive monotonic constraint in a default model. This means that the algorithm will only be allowed to predict that higher values of DTI result in higher predictions of default.

26. See Machine Learning and FICO Scores, available at https://www.fico.com/en/resource-access/download/6559.

27. See FICO Falcon X at https://www.fico.com/en/products/fico-falcon-x.

28. See endnote 6.

29. This could occur because race is often correlated with predictive factors that are unmeasurable or unavailable and are therefore not included in models.

30. See endnote 14.

31. Rudin, C. (2019), "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, 1(5), 206-215; Chapman, P. et al. "CRISP-DM 1.0—Step-by-Step Data Mining Guide", (SPSS, 2000); Agrawal, D. et al., "Challenges and Opportunities with Big Data: A White Paper Prepared for the Computing Community Consortium Committee of the Computing Research Association" (CCC, 2012), available at http://cra.org/ccc/resources/ ccc-led-whitepapers.

32. Risk patterns found within CRA data have remained quite stable through the years. For example, while the weighting has shifted somewhat over time, we understand that payment history and debt/utilization were the top two drivers of FICO Score from the earliest releases of the model and remain so today.

33. See Machine Learning and FICO Scores, available at https://bit.ly/30sk9rA.

34. See Fahner, G. (2018), "Developing Transparent Credit Risk Scorecards More Effectively: An Explainable Artificial Intelligence Approach", available at https://bit.ly/3ckTqQq.