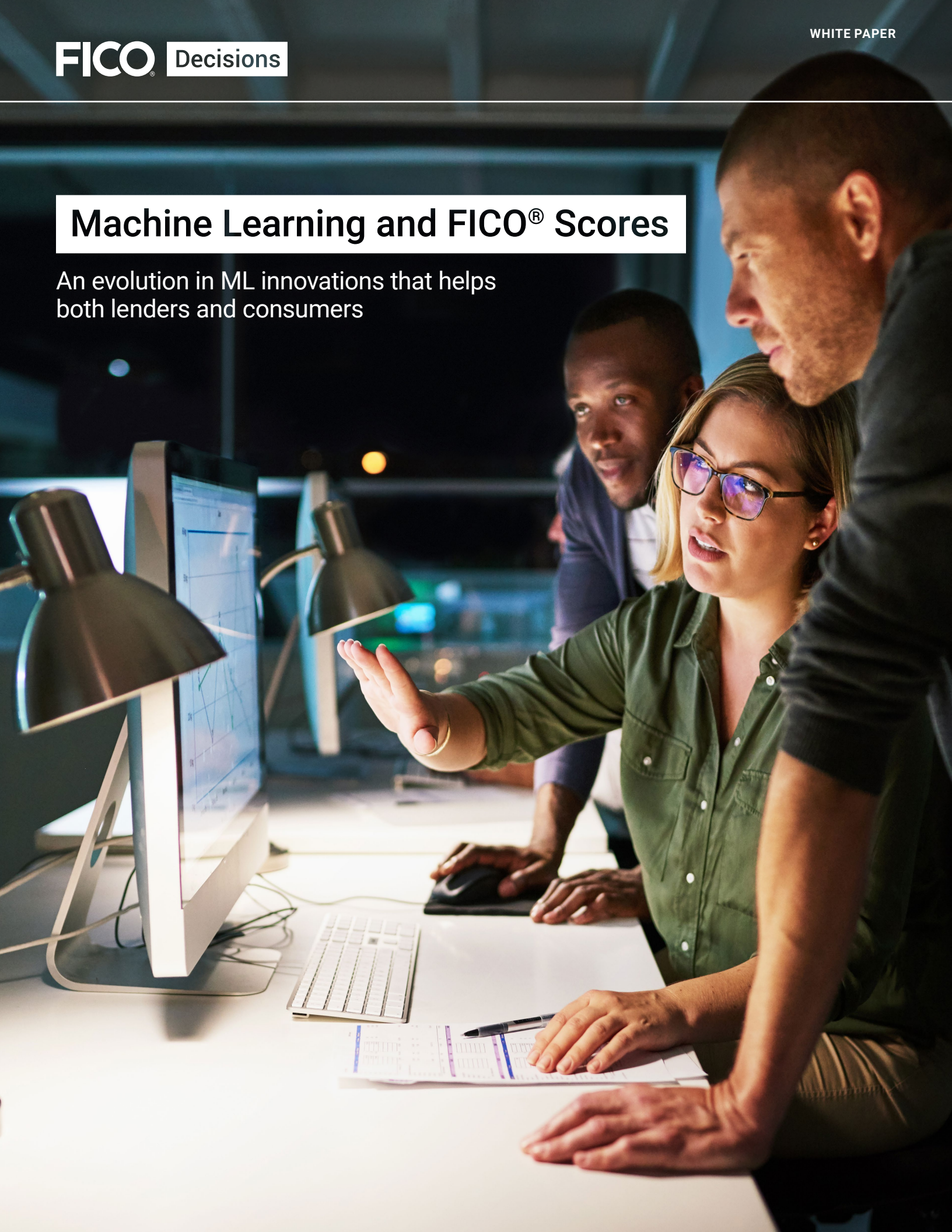
FICO Decisions

WHITE PAPER

Machine Learning and FICO® Scores

An evolution in ML innovations that helps both lenders and consumers



I | Introduction

A technology's journey from the lab to the field

Every day, new technology innovations are hatched in incubators, corporate research labs, universities and government institutions. But no matter where it originates, successful new technology follows a predictable evolution, migrating over time from nascence to mainstream maturity. As it should; immature technologies that are unleashed too quickly for widespread use — Google Glass, hoverboards, *et al.* — can have dangerous repercussions.

In the world of credit risk assessment, machine learning (ML) is one of these technologies. Thirty years ago, FICO began using early ML techniques in a lab environment; in the decades since, we have finely honed our ML expertise, which is necessary to leverage machine learning effectively and safely for applications in the field.

FICO continues to invest significantly to evolve the latest machine learning techniques and bring new innovations to FICO® Scores in global markets. Most recently we have incorporated ML into the FICO® Risk and Affordability Decision Suite for the UK market, which includes two new models, **a FICO® Customer Management Score and a FICO® Balance Change Sensitivity Index**. The latter of these, advances the predictive power of analytic solutions from correlation toward causation –– a major breakthrough. (See Section 5.)

ML technology cannot overcome a lack of data

There's a lot to marvel about the effectiveness of ML, but the technology is only as smart as the data it consumes. In the last decade or so, new players in the credit scoring market have promoted models built entirely or predominantly with machine learning as "more modern," innovative and effective bases for fair, inclusive credit decisions, particularly for underbanked and "unscorable" populations. FICO believes these assertions are overstatements, as no ML technique alone can overcome the fundamental lack of credit data available for these consumers.

Furthermore, overreliance on "ML-only" models can actually obscure risks and shortchange consumers by picking up harmful biases and behaving counterintuitively. Such models could underestimate default risk or deny consumers improvements to their credit scores as they lower their debt. This lack of explainability makes "black box," ML-only models difficult to operationalize at any scale and, in turn, unpalatable to lenders and consumers, particularly those who are inexplicably denied credit. Weak analytic accountability thus creates opportunities for market confusion, lender losses and consumer exploitation.

FICO's development approach enhances new technologies such as the latest machine learning with decades of domain expertise in building credit risk scoring models that are fair to all consumers, including unscorable populations, and withstand regulatory and lender scrutiny.

This paper helps a business audience to understand how FICO assesses ML techniques for possible inclusion in FICO® Score models. It describes the results of rigorous testing of ML-only credit scoring models against the FICO® Score, conducted by top FICO experts in ML. In doing so, this paper illustrates why the path to future innovations in credit scoring is a journey, not a race.

2 | Compare & Contrast

Machine learning models and the FICO® Score

Although both the latest machine learning algorithms and the FICO® Score analytic model can produce a credit risk score, their underlying technology is very different. The ML-only techniques we tested to develop credit scoring models include multilayer neural networks[1] and gradient-boosted decision trees[2].

Building a forest from shallow trees

Figure 1 illustrates how a typical ML-only process uses training data containing *predictors* and the *outcomes* the model is trying to predict. For this exercise, the outcome is whether a consumer will miss payments on credit obligations during a time period after the predictors were observed; the predictors are composed of thousands of credit bureau characteristics developed by FICO over 25 years of development of the FICO® Score.

[1] For more information visit: http://www.fico.com/en/predictive-analytics/analytic-technologies/neural-networks

[2] For a description of stochastic gradient boosting see Jerome H. Friedman, March 26, 1999. https://statweb.stanford.edu/~jhf/ftp/stobst.pdf

Anatomy of a Machine Learning Model

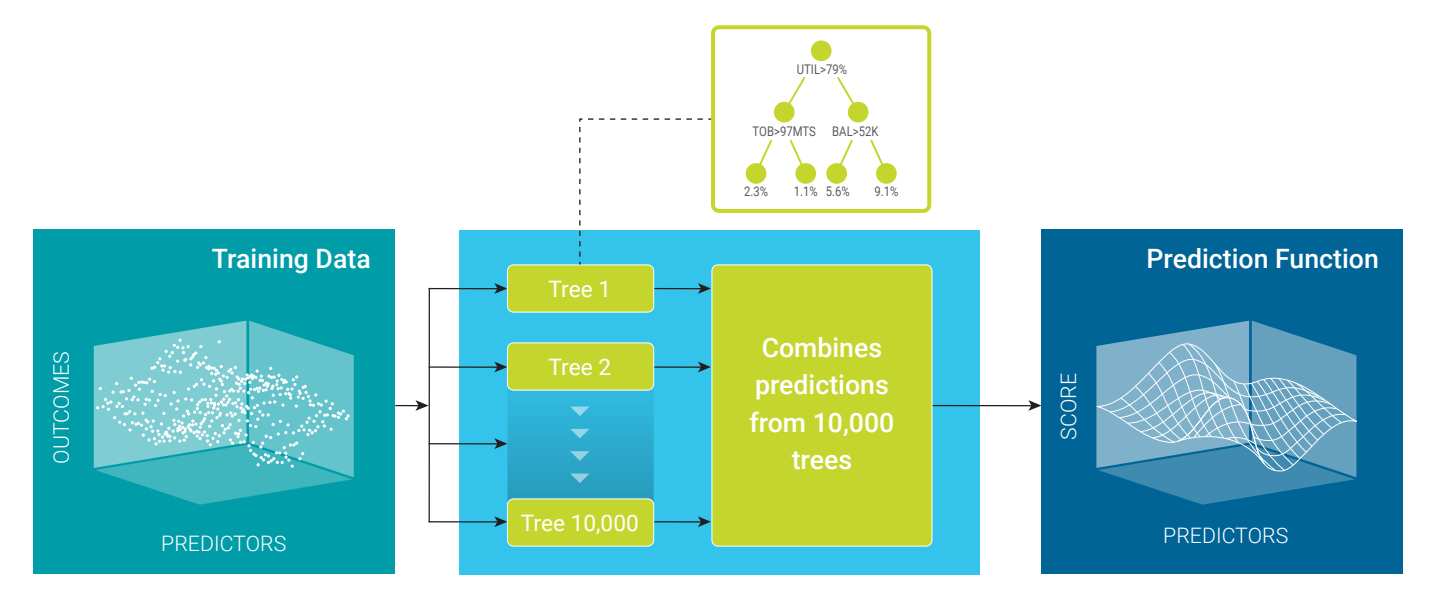


Figure 1: The gradient boosting approach uses training data to generate thousands of trees, which are combined to produce a predictive credit risk score.

The center portion of Figure 1 illustrates how the data is used to build thousands of shallow trees that segment the population. For example, the green dot shallow tree segments the population into groups that are using more than 79% of their available credit, or less; subsequent branches of the tree address the length of credit history and total debt balances. This process is repeated thousands of times leading to better and better predictions through gradient boosting; the resulting trees are combined to produce the output: a machine learning-driven credit score.

Because a gradient-boosted machine learning approach produces a credit risk scoring model composed of thousands of trees, it's a challenge to determine exactly which variables drive particular predictive outcomes and how. This challenge can be partially addressed by using simulation techniques such as Partial Dependence Plots[3] to gain insights into the input-output relationship of the resulting model.

On the positive side, gradient boosting captures nuanced patterns in the data automatically and effectively, which other model-building techniques not based on the latest ML (e.g. logistic regression) don't automatically do.[4]

Illustrative FICO® Score Scorecard

Category	Characteristics	Attributes	Points
Payment History	Number of months since the most recent serious delinquency	No serious delinquency	75
		0 – 5	10
		6 – 11	15
		12 – 23	25
		24+	55
Outstanding Debt	Overall utilization on revolving trades	No revolving trades	30
		Under 6%	65
		7 – 19%	50
		20 – 49%	45
		50 – 89%	25
		90% or more	15
Credit History Length	Number of months in file	Below 12	12
		12 – 23	35
		24 – 47	60
		48 or more	75
Pursuit of New Credit	Number of inquiries in the last 6 months	0	70
		1	60
		2	45
		3	25
		4+	20
Credit Mix	Number of bankcard trade lines	0	15
		1	25
		2	55
		3	60
		4+	50

Note: Graphic for illustrative use only

Figure 2: A typical FICO® Scorecard contains up to 20 variables in five credit risk categories.

FICO® Scores are produced by a system of segmented scorecard models

Unlike credit score models built solely or predominantly with ML, for which the output can be a blend of thousands of data-driven models, FICO uses the FICO® Model Builder Scorecard Module[5] technology to produce a system of segmented scorecard models that are:

- **Engineerable:** Constraints can be applied to test and refine each scorecard to ensure palatability and to overcome potential data weaknesses.
- **Transparent:** How the variables combine with each other to impact the score is very clear, and explainable. Figure 2 illustrates the explainability of one variable from each of the five key categories that compose the FICO® Score; a typical FICO Scorecard contains between 12 and 20 variables.

[3] For more information on Partial Dependence Plots see: http://events.fico.com/Macine-Learning-and-Human-Expertise (slides #36-40)

[4] Not without considerable manual effort to develop model enhancements such as complex characteristics capturing nonlinear transformations, interactions between raw variables and model segmentation.

[5] For more see white paper: http://www.fico.com/en/latest-thinking/white-papers/introduction-to-model-builder-scorecard

FICO® Score 9 utilizes 13 different credit risk scorecards tuned to distinctly different population segments, such as consumers with:

- A short credit history
- A long credit history
- A credit history with past major blemishes
- And more than a dozen additional segments

A multi-scorecard approach allows FICO to capture nuances in risk patterns and data interrelationships, meeting the same desirable criterion that is typically considered a machine learning strength.

The explainability challenge in credit risk scoring

In markets where credit risk scoring models are regulated and scrutinized, there is a strong requirement for the models, and the credit decisions derived from them, to be explainable. The impact each variable has on the credit score must be traceable (transparent), clearly explained and palatable (understandable and acceptable) to lenders, regulators and consumers.

These explainability and palatability requirements are guaranteed to be met by the current FICO® Score model construct. In contrast, ML-only models can be difficult to explain, requiring considerable simulation effort to even approximate how the models compute their scores. Even if explanations can be computed, they may not be palatable in all cases.

3 | Research In Action

Applying ML to challenge the FICO® Score

In FICO's continued efforts to assess whether the latest ML technologies yield performance improvements over FICO® Scores calculated via time-tested scorecard models, FICO researchers explored the different models' tradeoffs between:

- **Performance:** The efficacy in identifying individuals of acceptable credit risk
- **Palatability:** The acceptable explainability of the score, in order to pinpoint the impact of specific risk factors on a credit score.

FICO's research team found that building a gradient-boosted decision tree scoring model analogous to the FICO® Score took only 40 resource-hours, compared to the roughly 800 resource-hours typically required to build the scorecards that compose a FICO® Score model.

The relative ease of developing ML models lowers the barriers of entry for more ML-only model providers––which can produce market confusion, in contrast to the time-tested, regulatory compliant and explainable FICO® Score.

How the models were tested

The analytic models — FICO's own FICO® Score model and its ML-only counterpart — were A/B tested against the same dataset.

- **A/B testing** is an "apples-to-apples" comparison method in which the same data and predictors, as utilized in today's FICO® Score analytic model, are used to train an ML-only model. This allows the effects of each scoring technology to be isolated and identified.
- **The dataset** is the FICO® Score 9 Development Dataset, a nationally representative sample of millions of credit files.

As two representatives of ML-only models, FICO's researchers tested neural networks and stochastic gradient boosting (SGB).

ML-only models yield very marginal predictive improvement

The A/B testing revealed that ML-only models offer only very small predictive improvements. Figure 3 shows the nominal differences between the FICO® Score model and the ML-only models in two key credit risk performance metrics, the Kolmogorov Smirnov (KS)[6] statistic and the Receiver Operating Characteristic (ROC)[7]. When compared to the FICO® Score 9 model, the best ML-only models produced a modest predictive lift of less than 2% relative improvement in the KS metric. Likewise, the ROC curves of the FICO® Score and ML-only models are quite close.

[6] This statistic is the maximum difference between the cumulative distributions of non-defaulters and defaulters. A zero value indicates that the credit score fails to differentiate between defaulters and non-defaulters; a value equal to 100 indicates that the credit score perfectly differentiates defaulters from non-defaulters

[7] The ROC curve is a plot of the true positive rate (percentage of "bad" applicants rejected, e.g. defaulters) versus the false positive rate (percentage of "good," paying applicants rejected) found over a set of predictions. Associated with the curve is the ROC metric which measures the area under the curve. A value of 0.5 indicates that the score performs no better than random; a value of 1 indicates that the score perfectly differentiates defaulters from non-defaulters.

FICO® Score R&D

ML Lift Over Scorecard Approach is Measurable, but Modest

Outcome measure: performance on bankcard accounts over 24 months (bad = 90+ days past due)

Score	Metric	Value
FICO® Score 9	ROC	.901
	KS	63.2
SGB-Based	ROC	.905
	KS	64.3
Neural Net-Based	ROC	.901
	KS	63.5

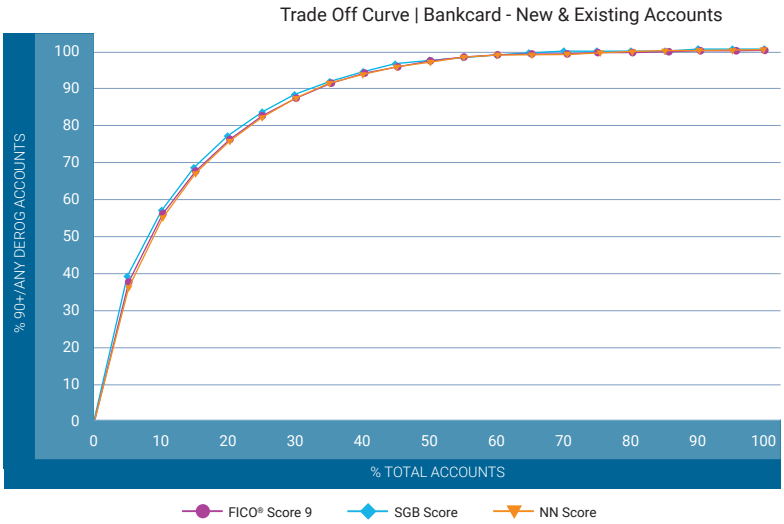


Figure 3: FICO researchers tested the efficacy of the FICO® Score model and two variations of a ML-only credit scoring model. All models were developed to predict defaults (payments more than 90 days past due) on all credit accounts using the same training data and their performance was evaluated on an independent test data set.

The score performance metrics and the trade-off curve illustrate the measurable but modest differences in the predictive power of the ML-only credit risk models and the FICO® Score model.

Testing ML-only against FICO® unscorable[8] populations

The next step in FICO's evaluation tested both the ML-only model and FICO's Scorecard approach on unscorable populations. In particular, we investigated whether the ML-only model may be able to squeeze out additional predictive information from the sparse credit data that is available on unscorable files. Unscorables' high rate of missing performance on loan repayments also raises the specter of selection bias, a major challenge that needs to be addressed to create a reliable model.

Using a sample of millions of unscorable records, FICO compared the predictive power of credit bureau-based "research scores" built via scorecard technology to those solely based on stochastic gradient boosting. Figure 4 contains the representative results: Compared to the scorecard model, the ML-only model produced an approximate 5% improvement of KS values and Gini[9] on the "derogatory info in credit file" segment. However, the scorecard-based model was engineered for palatability, while the SGB-based model was unconstrained.

[8] For more see white paper: http://www.fico.com/en/latest-thinking/white-papers/can-alternative-data-expand-credit-access

[9] Gini is a statistical measure of the degree of variation or inequality represented in a set of values, used especially in analyzing income inequality.

Importantly, when palatability constraints were removed, the scorecard model yielded nearly identical levels of ROC, KS and Gini.

Figure 4: In testing both a scorecard model and an ML-only model against unscorable consumer credit files with derogatory information, removing palatability constraints from the scorecard model produced predictive performance nearly identical to the ML-only model.

Performance of "Research Score" on Unscorable Credit Files with Derogatory Info

	ROC	KS	Gini
Scorecard-Based	0.639	19.7	.278
SGB-Based	0.647	20.8	.294
Scorecard-Based (no palatability constraints)	0.645	20.6	.291

When palatability constraints were removed, scorecard model yielded nearly identical ROC/KS/Gini

These results (note the much lower KS figures observed relative to the scoreable population represented in figure 3) indicate that all models struggle similarly when there is only sparse credit bureau information available. Any predictive lift attributable to the ML-only approach appears to be derived from sacrificing palatability for a modest improvement in model performance.

Addressing the selection bias challenge

Selection bias is a fundamental data problem with ML-only models because they are trained and calibrated only on "cherry-picked" cases, specifically, those posing better credit risk. ML-only models will therefore likely underestimate the true default odds for previously rejected consumers.

Selection Bias and Extrapolation Risk

Figure 5: ML-only credit risk models excel at fitting to training data but extrapolate unreliably across the truncation area.

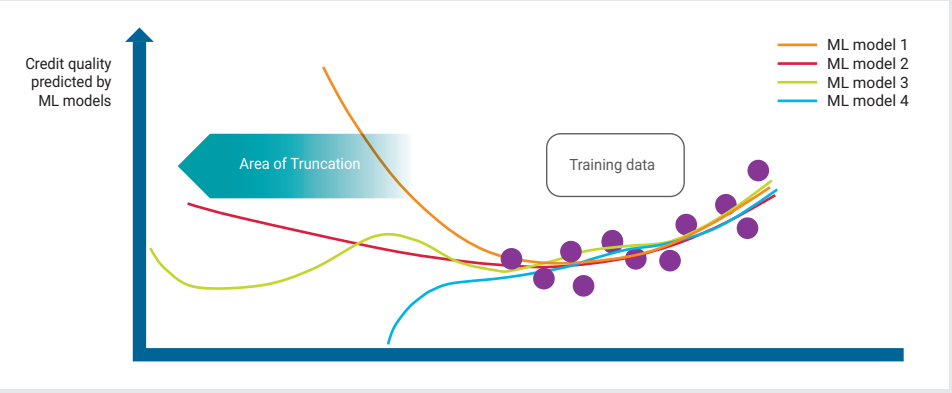


Figure 5 shows how pure-ML credit risk models excel at fitting to training data, illustrated by the performance of four different ML-only models. Here, the models agree that credit quality increases along the dimension of the training data plotted from left to right. However, the models extrapolate in idiosyncratic, uncontrolled, and unexplained ways across the truncation area to the left side of the plot. Therefore, the predictions of ML-only outside of the selection-biased training data are inconsistent, unreliable and cannot be trusted — a fact that has significant implications in production lending environments.

Human expertise is required

In sum, ML-only models don't provide a cure for data limitations, nor will they alert model developers and score users about masked risks.

Because they are unconstrained and, for large parts of the unscorable population, lack an important dependent variable (subsequent payment performance), **ML-only credit risk models are not equipped to counteract the significant selection biases due to truncation and cherry-picking that exist in unscorable populations.** This can result in biased predictions, erroneous credit granting decisions, and, in turn, negative customer experiences and lenders' failure to comply with fair lending regulations.

ML-only models adversely affect palatability

Paying down credit card debt improves credit scores, an axiom in the world of credit risk management. As part of our research, we leveraged a FICO® Score Simulator to quantify how paying off specific amounts of debt impacted the credit scores in this study.

Figure 6 captures the consumer palatability of the FICO® Score model as compared to the ML-only model. The unconstrained ML-only model built via SGB technology would result in 9.2% of consumer records receiving a *lower* score after debt was paid off (all other factors held equal). This effect is highly unpalatable; consumers and lenders in high-stakes, high-stress credit situations, such as applying for a home mortgage, would be confounded by such a diametric deviation from their mutual expectations.

ML-only models may deliver counter-intuitive, unpalatable results

Figure 6: In an ML-only credit risk model, positive credit action led to a lower score nearly 10% of the time. Zero percent of accounts scored via FICO® Score showed counterintuitive results.

Score Used	Result of Simulation Analysis - Paying Down Credit Card Debt
FICO® Score	**0%** of consumer records experienced a decrease in score as a result of this positive credit behavior
SGB-Based Model	**9.2%** of consumer records experienced a decrease in score

Separately and together, these test results provide evidence that ML-only credit risk models are unsuitable as the primary determinant of credit worthiness in mortgage and other types of lending. The ability to impose palatability constraints within the FICO® Score provided by scorecard technology ensures that counter-intuitive results are minimized, greatly improving the consumer and lender experience.

4 | Field Use

Bringing ML power to the field safely and effectively

As detailed above, FICO's research reveals that ML-only credit scoring models are very effective at capturing data's predictive content, making them highly effective for research and discovery work in a laboratory environment. But ML-only models have explainability and palatability shortcomings, rendering them imprudent to directly deploy into the field.

In comparison, FICO® Scores based on scorecards are easy to explain, a requirement for field use. But they require significantly more resource-hours for development. This presents a quandary when a new credit risk score must be developed that has to be highly predictive, explainable and palatable — and brought to market quickly.

FICO safely speeds ML innovation to market

The solution is to combine the strengths of ML-only models (discovering subtle predictive patterns in the data) with the strength of multi-scorecard models (highly predictive and easy to explain). The FICO Scores team executes this approach using a two-phase development strategy:

- Develop the best ML-only model quickly in the lab, using inherent highly automated processes.
- Closely approximate the best ML-only model by a system of segmented scorecards, also a highly automated process. Domain experts remain in control, imposing constraints on the scorecards to ensure explainability and palatability.

FICO recently used this approach to develop a new FICO® Customer Management Score in the UK, quickly deploying the resulting model into broad field use. This score is a component of a newly offered managed-service RegTech solution, the FICO® Risk and Affordability Decision Suite, powered by Equifax®[10] and is fully explainable and palatable, with predictive power that is close to that of the ML-only model.

Machine learning also helps to address challenges beyond assessing credit risk: FICO is using ML to understand consumers' "affordability risk" — the potential that lenders could induce financial distress in their customers through lending decisions. This analytic leap from correlation to causation benefits both from the effectiveness of ML to capture subtle relationships, as well as from the aforementioned two-phase development strategy to ensure explainability and palatability of the resulting model.

[10] For more information about FICO® Risk and Affordability Decision Suite, visit http://www.fico.com/en/node/8140?file=14027

5 | Conclusion

It's a journey - not a race - to the future

ML augments human expertise but doesn't replace it

At FICO, our mission is to innovate new tools that enable lenders to safely expand consumer access to credit and fuel economic growth. Machine learning is an exciting technology that we have used for more than 30 years to enhance FICO® Scores globally. We remain at the forefront of exploring how ML can be applied to quantify credit risk, and identify key drivers of default, fueling new discoveries in the analytic leap from correlation to causation.

But our research also reveals the fallibility and adverse potential of ML-only scoring models:

1. ML-only models are not a cure-all for a lack of data.
2. ML-only models can produce potentially biased predictions and underestimate default rates in traditionally unscorable populations.
3. ML has limited predictive upside over a well-constructed system of scorecards.
4. ML-only models potentially lack transparency and palatability.

Unleashing ML-only models into the broad lending market would almost certainly usher in systemic risk, market confusion, and lack of transparency for consumers. FICO's long-standing philosophy is that innovations in ML must be combined with domain expertise and should be complemented with the provision of relevant new data sources, an approach that drives the time-proven safety, soundness and innovation of FICO® Scores. The path to the future is a journey, not a race.

Lenders interested in learning more can contact us at ficoscoreinfo@fico.com.

To keep tabs on the latest FICO research on scoring best practices and credit risk trends, visit the FICO Blog.

FICO

FOR MORE INFORMATION
www.fico.com
www.fico.com/blogs

NORTH AMERICA
+1 888 342 6336
info@fico.com

LATIN AMERICA & CARIBBEAN
+55 11 5189 8267
LAC_info@fico.com

EUROPE, MIDDLE EAST & AFRICA
+44 (0) 207 940 8718
emeainfo@fico.com

ASIA PACIFIC
+65 6422 7700
infoasia@fico.com

4611WP_EN 09/18