Fair Lending in the Brave New World of Big Data

By Marsha J. Courchane, Ph.D. and David M. Skanderson, Ph.D.

May 2017

A version of this paper appeared in ABA Bank Compliance, May/June 2017

The burgeoning new world of online "marketplace" lending holds the promise of broader access to credit for consumers and businesses with reduced cost and more competition. However, cutting-edge technology, data sources and credit modeling methods also pose potential for fair lending risk – notwithstanding the fair lending benefits of automated decision-making. We provide an overview of marketplace lending, the "Big Data" and machine-learning technologies increasingly being used by both online and traditional lenders, and key considerations in managing fair lending risk in this brave new world.

Exhibit 1

Some typical characteristics of marketplace lenders:

- Operate exclusively online
- Niche product/market focus
- High degree of automation with little or no scope for system overrides or policy exceptions
- Proprietary and innovative predictive models for marketing, credit scoring, fraud detection and pricing
- Use of non-traditional data sources
- Rapid pace of change in decision criteria and scoring models

Regulators have taken a keen interest in the sector, as well. The federal financial regulatory agencies appear to be attempting to adopt a fairly constructive and balanced approach to marketplace lending, Big Data and machine learning. The U.S. Treasury Department, Federal Deposit Insurance Corporation, Consumer Financial Protection Bureau (Bureau) and Federal Trade Commission have each published reports on these topics that have recognized both the potential benefits and risks to consumers and small businesses, and in February of this year the Bureau issued a Request for Information seeking to learn more about the topic.[1]

What is Marketplace Lending?

We've all heard the term 'marketplace lending', but what does it mean and who are these new competitors and cohorts? Generally the term refers to technology-focused online lending platforms which connect sources of capital (individual and institutional investors) to users of capital (consumers and small businesses). The original peer-to-peer lending platforms from which the sector emerged allowed individual investors to fund loans to individual consumers, with the platform serving as matchmaker, screening mechanism and servicer. The sector grew rapidly with funding from venture capitalists, hedge funds, and institutional investors as many banks cut back their consumer and small business lending after the financial crisis. More recently, equity and debt financing, direct bank funding and securitization have played an increasing role in funding online marketplace loans. While some of these lenders compete against banks, many financial institutions are partnering with or investing in these new lenders so it is important to understand them.

Marketplace lenders have filled unmet demand for credit by leveraging new technology to compete directly with higher-cost and less convenient traditional lenders, and by providing opportunities for capital in search of higher returns in a low interest rate environment. Marketplace lenders generally offer more flexible and innovative products, more favorable pricing to consumers, quicker decisions and more efficient service. Some of the typical distinguishing characteristics of most marketplace lenders are listed in Exhibit 1.

Exhibit 2

Examples of marketplace lending products and some of the lenders:

- Unsecured personal loans & credit lines: Avant, Lending Club, LoanDepot, Prosper
- Education lending: Common Bond, Earnest, SoFi
- Small business loans & credit lines: Funding Circle, Kabbage, OnDeck
- Small business receivables finance: Blue Vine, Fundbox
- Point-of-sale finance: Affirm, Bread
- Vehicle secured: AutoFi, DriverUp, LendKey
- Real estate secured: LendingHome, Realty Mogul, SoFi

Many marketplace lenders have focused on niche markets that had been overlooked historically by banks or that had been the province of higher-priced finance companies and payday lenders. Some of the lenders have focused on refinancing and consolidating higher-interest-rate debt to lower rates, including student loans and credit card debt (see Exhibit 2 for the most common types of products and some typical players). To some extent, all have focused on using cutting-edge technology to identify and acquire creditworthy consumers and to originate loans at low cost. While they still rely heavily on traditional direct mail marketing and advertising media, they also mine social media, online ad placement and lead generators for prospects.

The role of Big Data

In consumer and small business finance, "Big Data" usually refers to the practice of collecting and combining large amounts of data about potential borrowers from many diverse data sources – well beyond the three national credit reporting agencies – and using computationally intensive processes to discover patterns and interrelationships in the data that help to understand borrower habits and predict credit behavior. Diverse data sources can offer improved insights into consumer behavior. Many of the marketplace lenders are heavy users of Big Data, although banks are also moving in this direction.

The data sources may include spending and shopping behavior, bank account activity, sources of credit used, data from alternative credit reporting agencies, online and social media activity and various other sources such as those listed in Exhibit 3.

Note that not all marketplace lenders rely on alternative data sources, however. Many rely on traditional credit bureau scores (such as FICO and Vantage scores) and ability-to-repay measures (such as debt-to-income ratios or disposable income measures); while others use a combination of traditional and non-traditional credit history, behavioral and other attributes; and still others exclusively use proprietary credit models and decision systems. Some use no credit score information at all.

Exhibit 3

Examples of non-traditional data that may be used in marketing, credit decisions, identify verification and fraud risk detection:

- Payday and non-prime loan information
- Check cashing services
- Rent-to-own transactions
- Mobile phone account openings & payments
- Utility accounts & payments
- Property-related public records
- Rental applications and evictions
- Trends in income
- Local unemployment rates, property values, income levels, economic growth
- Bank account transaction data (timing of account openings, number of accounts, pay frequencies, returned items, debit transactions)
- Social media and web surfing data
- Address stability
- Number and age of e-mail addresses
- Education (degrees earned, school, field of study, graduation year, student loan cohort default rates)
- Profession or job function
- Direct link to borrower bank accounts, small business accounting software or merchant payment processing systems

The role of machine learning

In marketplace lending, complex new "machine learning" models are augmenting or replacing traditional credit scoring methods. The traditional banking market is migrating this direction too. Machine learning uses computers to create analytical models on an automated basis, and to make decisions without being programmed with a specific set of decision criteria. Based on the available data, general model structure and business objectives defined by the model developer, a computer determines which data elements to use in predicting credit behavior and which consumers to solicit or applicants to approve. Familiar everyday examples of machine learning from outside of the financial world are the highly personalized advertising on Facebook and product recommendations on Amazon.

Traditionally, credit scoring uses statistical analysis to derive a fixed formula based on a defined set of credit history attributes, each of which is assigned a fixed number of points resulting in a numeric credit score. Scoring models are usually adjusted or replaced over time as predictive performance degrades, but usually no more often than annually.

By contrast, machine learning models are flexible algorithms that grow and change with exposure to new data. Such modeling methods can be adept at crunching through large volumes of data to identify characteristics and their interrelationships that help to predict credit behavior. Machine learning methods allow the model to build and update itself. The automated model determines which variables are useful and how to combine them to best predict behavior based on the latest data available. The model determines which applicants should be approved based on targets for business objectives (e.g., charge-off rates or profitability), without necessarily computing a numeric credit score. In some cases, machine learning may combine the results of multiple models to increase predictive power. Finally, machine learning methods tend to select attributes and combinations of attributes based purely on the strength of their correlations to credit outcomes. Less emphasis is placed on whether there are logical economic or behavioral reasons underlying those correlations..

A key motivation for machine learning is the desire to identify and exploit subtle and difficult to observe relationships among disparate data elements from many different sources that can be combined to better predict consumer behavior. As a simple example, there may be a distinct difference in default risk between a consumer with multiple recent delinquencies on a single credit account and several recent credit card inquiries who has also recently applied for a payday loan, compared to a consumer who is otherwise the same but has delinquencies across multiple accounts. It would be prohibitively time consuming or impossible for a human analyst to evaluate all possible interrelationships among all available data elements to identify the characteristics that best predict default. Machine learning techniques, however, can allow an analyst to consider an arbitrary number of complex interrelationships among hundreds or thousands of variables.

Beyond that, because the process of building the model is automated, it can be updated frequently as new data on actual loan performance becomes available. This is especially valuable for new products and new lenders, for which loan performance experience is very limited.

Fair lending benefits and risks

Big Data and new modeling approaches have the potential to provide new insights into consumer behavior that could improve profitability for lenders and broaden credit access for consumers. Alternative data sources and modeling methods could allow lenders to better serve consumer segments that historically have been underserved, such as consumers who are unbanked, have low or moderate incomes, do not use traditional credit products, are self-employed or have little established credit history.

For example, a 2015 study by the Bureau estimated that about 15% of Blacks and Hispanics are "credit invisible" – meaning that they have no records at the national credit reporting bureaus – compared to about 9% of Whites and Asians.[2] The study also found that a further 13% of Blacks and 12% of Hispanics have credit bureau records that cannot be assigned a traditional credit score because of insufficient credit history or insufficient recent credit activity, compared to about 7% of Whites and Asians. Even without understanding the assumptions, data attributes, or motivations in the study, the results suggest that mining alternative data sources for information about consumer payment behavior or risk characteristics could potentially broaden access to credit for minority consumers.

Easy and low-cost access through an online platform generally results in faster credit decisions and funding, reduced shopping costs, reduced geographic boundaries, increased choice and flexibility for consumers seeking credit and provides opportunities to build good credit management habits. Also, the automation of credit application and decision processes reduces the risk of disparate treatment on a prohibited basis that can arise in manual or judgmental processes – assuming the inputs to the automated decision are not problematic.

Fair lending risk can still arise with automated and machine learning processes and fair lending risk management becomes more challenging with machine learning and data analytics. It's probably safe to say that there is not a full appreciation among credit risk specialists of how fair lending risk may arise in automated, model-driven processes. Modelers are likely to say, "We don't discriminate. Our models don't consider prohibited factors." While that's a big step in the right direction, the risk of disparate impact (of particular attributes alone or in combination with other attributes) may not receive sufficient attention, and not all aspects of the credit process may be evaluated for fair lending risk.

There are various potential sources of fair lending risk that should be considered in the use of alternative data and automated decision processes. First of all, the Equal Credit Opportunity Act (ECOA) and Regulation B prohibit lenders from discriminating on the basis of prohibited characteristics in <u>any</u> aspect of a credit transaction. This means that the full credit lifecycle must be evaluated for fair lending risk, including marketing, underwriting, fraud risk detection, setting terms and conditions (pricing, credit line/limit determination, etc.), servicing and collections. Each of these stages in the process may involve different data sources, decision criteria and models with different fair lending risk potential, and each should be evaluated.

Next, the risk of a disparate impact on a prohibited basis should be evaluated. Ostensibly neutral variables that predict credit behavior may nevertheless present disparate impact risk if they are so highly correlated with a legally protected characteristic that they effectively act as a substitute for that characteristic.

Some alternative data elements that may be used in credit models and decisions have well recognized correlations with prohibited factors, posing disparate impact risk. For example, geographic location, use of banking services, educational attainment, college or university attended and use of nonprime credit tend to be correlated with race and ethnicity. Unlike credit history data, which has long been accepted by regulatory agencies as having a legitimate business justification notwithstanding its correlations with prohibited bases, alternative credit attributes have yet to gain widespread acceptance and some are viewed with suspicion. If such factors are used in credit decisions, lenders should be diligent in developing rigorous evidence of their business justification and in evaluating them for potential fair lending impacts.

Predictive variables based on aggregated information should be given particular attention. For example, the average credit risk of people residing in a given geographic area, local economic factors (e.g., local unemployment rates or property appreciation rates), or indicators of a consumer's memberships or affiliations can pose a fair lending risk because they tend to treat large groups of consumers similarly regardless of their individual risk characteristics and because the aggregated factors may be correlated with one or more prohibited bases. For example, a geographically based predictive variable would treat high-risk and low-risk consumer within a given geographic area as having the same risk (other things equal). If geographic location is correlated with race or ethnicity, then the geographically based variable may end up having an unjustifiable disparate impact even though it appears to help predict credit risk on average.

The use of alternative data sources and machine learning methods can also create or perpetuate biases if there are biases inherent in the data sources themselves. If data sources used are not representative of the population of consumers potentially qualified for the credit product of interest and/or systematically exclude certain segments of the population, they may tend to create feedback loops that perpetuate or reinforce historical biases. For example, reliance on data about banking behavior could tend to underrepresent certain minority groups, and reliance on data about social media or online shopping behavior could tend to underrepresent certain age groups.

Attention also should be given to the potential risk of redlining, reverse redlining or predatory lending. Such risks could arise from targeting credit products to such niche markets as higher-income or higher credit quality consumers, lower-income or nonprime consumers, consumers who are internet-savvy and communicate heavily through social media, or consumers who have a large "data footprint."

Managing the risk

Assessing, quantifying, and weighing fair lending risks of credit models and alternative data sources is a complex technical endeavor, and it is further complicated by the use of complex modeling methods such as machine learning. Nevertheless, compliance personnel can get a good sense of the potential fair lending risk by asking the right questions and evaluating whether necessary controls are in place.

First, we recommend scrutinizing the data sources and predictive variables being used as inputs to models and decision rules. Of course, no legally prohibited factors should be used with the possible exception of the age of the credit applicant. In the event that an age-split scoring system is used, it is important to ensure that it meets the "empirically derived, demonstrably and statistically sound" standard and other requirements of Regulation B. More broadly, it is important to evaluate the representativeness of the data used to develop a model, and whether any of the variables/attributes (or combinations of variables) in the model are likely to have strong correlations with a prohibited basis or otherwise might be controversial.

It is also important to evaluate the relevance of the variables in a model to the behavior or outcome the model is designed to predict. A consumer's past credit performance or current financial situation have direct intuitive relationships to future credit performance. However, data elements that appear to have predictive power but have no intuitive relationship to the credit behavior being predicted, should receive extra scrutiny. Such elements can be challenging to defend in the event that they create a disparate impact, and also can be challenging to explain to consumers in terms of adverse action reasons if they result in a denial decision. When potentially risky or questionable variables are encountered, it is important to evaluate how much they actually contribute to the predictive power and business objectives of the model, and to weigh those benefits against fair lending risk in deciding whether the variables should be used. A variety of statistical tools can be useful in evaluating the tradeoffs.

Second, it is important to ensure that models receive a rigorous statistical validation by a qualified, independent internal or external party to ensure the models are statistically sound and were developed according to generally accepted statistical methods. Statistical validity is an important line of defense against potential disparate impact claims. If a model or decision variable is found to have a disparate impact on a prohibited basis, it may still be permissible (i.e., not illegal discrimination) if its use is supported by a sufficient business justification. Statistical validation is aimed in part at confirming the evidence of that justification.

Third, it is important to ensure that a model's performance is regularly monitored over time. If the predictive power of a model quickly degrades over time, that is a sign the model might not have been statistically valid to begin with or that the correlations on which the model was originally based may have been idiosyncratic to the particular data sample or time period used to develop the model.

Fourth, it is important to identify and address any circumstances where the decisions of the automated system may be overridden and where there may be human touch-points in the process. For most of the marketplace lending processes the authors have observed, exceptions are very rare or non-existent, and the only human interaction in a loan application process might be in reviewing potential fraud risks and in requests for additional information from the applicant. Where there are opportunities for human intervention in the decision process, controls testing, quality control and statistical analysis can be used to diagnose the potential for fair lending risk.

Finally, retention of relevant documentation and data is critical to managing fair lending regulatory risk. Traditional banks know the importance of full documentation and retention, but this level of attention may not yet be in place for marketplace lenders. The data used to develop a model and documentation of the model development and validation processes should be retained because they are likely to be needed to perform fair lending testing and in the event it is necessary to defend against a discrimination claim. The data used in each credit decision should be retained for the same reasons. If application data is updated and overwritten over time (as sometimes occurs when the lender has an ongoing lending relationship with the consumer) it may be impossible to confirm in a retrospective review why a consumer was approved or denied, and thus to defend all of the credit decisions made.

Marketplace lending, machine learning, and Big Data offer important benefits to both those who supply and those who demand credit. For stable and responsible growth in this area to continue, responsible lenders will couple strong fair lending compliance oversight with their emphasis on revenue generation and credit risk management.

About the Financial Economics Practice at CRA

CRA's Financial Economics Practice provides economic and financial analysis and advice to financial institutions, financial regulators, and counsel representing financial institutions. Our experts are skilled in quantitative modeling and econometrics, particularly as applied to issues in credit and compliance risk in primary and secondary consumer lending markets. To learn more about the practice, visit **www.crai.com/financialeconomics**.

Contacts

Marsha Courchane, PhD
Vice President and Co-Practice Leader
Washington, DC
+1-202-662-3804
mcourchane@crai.com

David Skanderson, PhD
Vice President
Washington, DC
+1-202-662-3955
dskanderson@crai.com

Sources

[1] Consumer Financial Protection Bureau, "Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process," Docket No. CFPB-2017-0005, Federal Register, Vol. 82, No. 33, Tuesday, February 21, 2017, **https://www.gpo.gov/fdsys/pkg/FR-2017-02-21/pdf/2017-03361.pdf**.

Federal Deposit Insurance Corporation, "Marketplace Lending," Supervisory Insights Journal, Winter 2015, Vol. 12, Issue 2, **https://www.fdic.gov/regulations/examinations/supervisory/insights/siwin15/si_winter2015.pdf**.

U.S. Department of The Treasury, "Opportunities and Challenges in Online Marketplace Lending," May 10, 2016, **https://www.treasury.gov/connect/blog/Documents/Opportunities_and_Challenges_in_Online_Marketplace_Lending_white_paper.pdf**.

Federal Trade Commission, "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues (FTC Report)," January 2016, **https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf**.

Office of the Comptroller of the Currency, "Recommendations and Decisions for Implementing a Responsible Innovation Framework," October 2016, **https://www.occ.gov/topics/responsible-innovation/comments/recommendations-decisions-for-implementing-a-responsible-innovation-framework.pdf**.

[2] "Data Point: Credit Invisibles," Consumer Financial Protection Bureau, Office of Research, May 2015. **http://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf**.