Lessons Learned:
What We Would Hope to See When Using Machine Learning Models and When Evaluating Fairness

Dr. Marsha Courchane
Vice President & Practice Leader
Financial Economics
mcourchane@crai.com
202-662-3804

Dr. Adam Gailey
Principal
Charles River Associates
agailey@crai.com
202-662-3879

CRA Charles River Associates

Historical Perspective

- Four primary types of fair lending investigations:
  - Underwriting
  - Pricing
  - Redlining
  - Credit scoring/disparate impact (see Avery, Brevoort, Canner, *Real Estate Economics*, Dec 2012)
- Regulatory Trends
  - Mortgages were primary focus due to availability of race/ethnicity data
  - Around 2013, BISG applied (retrospectively) to indirect auto lending
    - BISG approaches have evolved over time (threshold or classification, continuous, Max)
    - Known biases exist – due to characteristics of analysis sample not aligning fully with characteristics of U.S. Census data on which the BISG proxy model relies
  - Fair lending in UW relied primarily on logistic models
  - Fair lending in pricing relied primarily on multivariate regression models
  - Redlining models relied on simple comparisons of lenders to some peer set (for example, similarly sized (50 – 200% of own volumes))

Effectiveness of Traditional Approaches

- There was regulatory consistency in the approach – even if thresholds for concern were unknown
- Underwriting
  - Odds ratios and marginal effects used to evaluate relative probability of denial of a member of a protected class
  - Manual file review nearly always needed to capture decisions made based on data not captured electronically
  - No regulatory guidance ever provided on the threshold of concern – was an odds ratio of 1.50 and below a "safe harbor?" or 1.20?
  - Very few referrals to DOJ based on UW
- Pricing
  - Models incorporated a variable to define the protected class and evaluated its coefficient for evidence of discriminatory pricing
    - Models could include one race group v. non-Hispanic white or include controls for each protected class in a single regression.
    - No regulatory guidance ever provided on the threshold of concern – was 5 basis points and below a safe harbor?
    - Key focus was on "discretionary" changes to price

Effectiveness of Traditional Approaches

- Redlining
  - Focus on particular MSAs where lender traditionally had 100 apps or originations (may now be 30 per year over 3 years)
  - Peer definitions add complexity to redlining examinations
  - Regulators want a simple peer definition (maybe volume screen)
  - Lenders want a peer definition that reflects their business
    - Depository v non-depository, Conventional v FHA, First lien v HELOC, Private banking
  - Concept of redlining has changed – few, if any, MSAs without several lenders doing business; much less reliance on physical branches
- Credit Scoring – there was a general belief that credit bureau data, and the resulting scores, satisfied the business necessity for disparate impact –
  - Traditional models used a few variables from the bureaus (often just the derived credit score) and were easily interpretable and transparent
    - These models generally monotonic
    - Easy to offer consumer adverse action reasons

Data, ML, and AI– Change is here to stay

FinTechs / Marketplace Lenders/ Large Banks/ Small Banks – where do we see changes

- Unsecured personal loans & credit lines
- Education lending
- Small business loans, credit lines and receivables financing
- Vehicle secured loans
- Real estate secured loans

Applications

- Underwriting – in many contexts
- Fraud Detection
- Verification of Income and Assets
- Pricing and/or Credit Line Assignments
- Servicing and Collections
- Portfolio Risk Assessment

The Transition to ML Models

- There is no regulatory safe harbor - financial regulatory agencies are studying this transition in various ways, but have not published any guidance upon which lenders can rely
  - OCC Office of Innovation, CFPB Compliance Assistance Sandbox, etc., FDIC webinars on “Banking on Data”
  - What method should be used for assessing differential outcomes vs business justification?
  - Automated LDA searches vs the modeler’s not having protected information
    - Cannot use protected class status in the development of the models (except for age in some circumstances)
    - This offers some benefits, but also means automated consideration of trade offs between fairness and credit outcomes cannot be done by “first line” model developers

Models and Modelers Have Changed

- Fundamental shift from linear regression or logit to much more complex models of many different forms
  - Neural networks, Tree based (gradient boost etc)
- More models used by more lenders (from large to small) and for more purposes (fraud, UW, marketing, etc).
- Data scientists / model developers (may) lack a deep (or any) understanding of the fair lending laws and traditions
- Substantial shift away from interpretability and explainability
  - Lots written on how to make the models more explainable/interpretable – but does not come close to the ease with which that naturally occurred when using linear regression models

Variable Inclusion

- Recent ML models, which use bureau data, generally do not justify the inclusion of the many (or even hundreds) of variables constructed from bureau data – rather they focus on model fit and predictive power more than the inclusion of any one variable or group of credit bureau variables
    - May require monotonicity, but not always
    - Techniques now developed that can lead to AANs
    - Does not mean models are easily interpretable or transparent
- The vast number of variables and their interactions mean that a single event (job loss, pandemic) may cause changes in a large number of these variables, so hard to see what the overall impact is
  – Less emphasis on having parsimonious models and more emphasis on letting the models learn and find interactions/correlations

Data – What is Being Added to Models? Pre-Processing? Validation?

- Social Media Data
- Bank Account Transaction Data (e.g. cash flow – from vendors such as Plaid)
- Accounting Software Data – to be used directly for UW
- Large data sets from open-source repositories or procured commercially from third party data aggregators
- Issues of redundancy, noise, overfitting all arise

Alternative	Traditional
Overdrafts in bank account	30/60/90 DPD
Current status of utility bills	Current status of tradelines
Major in college	Level of educ. attainment
Active invoices in accounting software	Dunn & Bradstreet rating

Potential Fair Lending Benefits and Risks

- **Benefits**
- Reduced discretion (to the extent it occurs) leads to lower potential for implicit bias or disparate treatment
- Increased access to credit for potential borrowers that can not qualify using traditional methods or data
  - See FinRegLab's working papers on access to credit (www.finreglab.org)
- **Risks**
- Explainability takes more effort and different techniques (not just p-values and r-squared)
- ML models may be susceptible to poor data quality (missing values)
- ML models may overfit (dump in the kitchen sink – e.g. ALL 900+ bureau fields)
- May be biases that are less easy to understand due to complicated interactions among many variables
- May be biases embedded in the bureau data

Assessment of Fair Lending Risk from Machine Learning / Alternative Data

- **Step** 1: Identify all of the models in use – whether developed internally or purchased from others
- **Step 2**: Determine which of the models have *the potential* to create differences in outcomes for current or potential customers
- **Step 3**: If there appears to be discrimination based on standard metrics (adverse impact ratio or standardized mean differences, for example), select the least discriminatory model that allows you to meet your business objectives (Traditional approach).
- **Step 4**: Dual optimization (use demographic features to decrease weighting) or Adversarial debiasing (Two competing models – one with and one not with demographic data).

Document Development Process – tradeoff between transparency and what some view as IP

Having a well developed approach to model risk management is key when using ML models. Some things to include:

- What is being predicted?
  - Is the outcome a business-justifiable outcome? Is there data to support that?
- Variables considered – which ones are kept? – Modelers should keep all development data
- Document correlations with intended outcomes
- Document the variables available to ML/AI system and any and all rationales for their inclusion
- How frequently do you update? New models? Or just tweeks? Document the chronology of any changes
- Ensure the outcome actually measures what it is purported to measure -- don't hide a credit model within a fraud model

Machine Learning Approaches Test 1000s of Models

- Machine learning often involves estimating and testing thousands of models
- Frequently many models have similar predictive ability – but may differ with respect to disparate impact
- Modelers typically select the "*Best Performing*" model based on predicting the outcome, robustness, etc.
- Could they consider alternative objectives that may predict similarly, but have less impact?
- There may be clear tradeoffs between predictive power and differential outcomes by protected class status

Key Areas of Potential Fair Lending Risk

- Does the model partially re-engineer societal bias?
- Does the model have specific factors that are likely to correlate highly with protected classes?
  - If so, are they acting as a proxy for different groups? Are they adding meaningfully to the model's predictive ability? Are there potentially less discriminatory alternatives?
- Does the model create the potential for redlining risk?
  - Geographic controls and neighborhood characteristics would not likely concern a modeler who lacks fair lending experience
  - Online marketing models may unintentionally target different audiences with different products, or not target some audiences to the same extent
- Almost all statistical testing outside of the mortgage market has to be done using proxies.
  - How good/bad are the proxies the regulators are using? In your own data, if you have mortgage data, you can test BISG for your footprint.

Assessing Model Fit – Various Approaches – but none that yet have any regulatory approval/discussion

- Confusion Matrix -- 2x2 matrix of Predicted (0,1) v True (0, 1)
  - Prediction vs Actual
  - Type I and Type II Error
  - Most other model fit metrics are based on this concept
- Area Under the Receiver Operating Characteristic Curve (ROC AUC)
  - (0.5 is low, 1 is high)
- Precision - # of positive class predictions that actually belong to the positive class (= True positives/(True Pos + False Pos))
- Recall - # of positive class predictions made out of all positive examples in the dataset (= True Pos/(True Pos + False Neg))
- F1 Score – balances precision and recall (poor = 0; best = 1)
- F2 – used when one class may be more costly than another – e.g. predicting a bad customer to be good (and approving loan) could be more costly than predicting a good customer to be bad (and denying the loan)
- K-S Statistic

Build Fairness in the Machine?

- Standard model minimizes risk
- Other models minimize risk, while penalizing models for creating different outcomes across protected classes
  - Adversarial neural networks
  - Fairness aware models
- Challenges with multiple protected classes – what if changes to model improve outcomes for Hispanics, but worsen outcomes for African Americans?
- Legal issues for first line development – cannot use demographics in model development
- How should you weight the fairness component?

Regulators Can and Should Help Allay Confusion

- Research the biases from applying proxy methodologies

  As ML models are used in so many areas, and only mortgage data allows collection of GMI data, this becomes ever more critical
- Update model risk management guidance
- Provide clear guidance as to what measures will be applied and accepted