A Practical Guide to Fairness Testing

For Legal & Compliance Teams

•••

Bell Analytics

April 2026

Disclaimer: We are not lawyers and cannot give regulatory advice. The following framework and case studies represent a data scientist's interpretation of reasonable standards and current stakeholder expectations. Best practices will evolve as additional regulatory guidance emerges.

Agenda

- Administrative Reminders
- Presentation
- Q&A (please use the “Chat” Feature)
- Closing Reminders

Administrative Reminders

- The Presentation
  - Available now at: https://cefli.org/webinars
- Post Event Communication
  - The presentation deck
  - A link to the recording
  - A “Certificate of Attendance” template
- Questions welcomed throughout the presentation (please use the “Chat” Feature)

CEFLI Premier Partners

CEFLI Affiliate Members

Gold

Silver

Bronze

CEFLI Antitrust Reminder

The Compliance and Ethics Forum for Life Insurers (CEFLI) is committed to adhering strictly to the letter and spirit of the antitrust laws. Meetings conducted under CEFLI's auspices are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CEFLI meetings be used as a means for competing companies or firms to reach any understanding -- expressed or implied -- which restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition. Accordingly, appropriate objection will be made to any presentation or colloquy that presents a risk from the standpoint of the antitrust laws.

A Practical Guide to Fairness Testing

For Legal & Compliance Teams

•••

Bell Analytics

April 2026

Disclaimer: We are not lawyers and cannot give regulatory advice. The following framework and case studies represent a data scientist's interpretation of reasonable standards and current stakeholder expectations. Best practices will evolve as additional regulatory guidance emerges.

30%
Of applicants are hired

25%
Of applicants are hired

Bell Analytics

30%
Of applicants are hired

25% → **83%**
Of applicants are hired Adverse impact ratio

30%
Of applications are approved

25%
Of applications are approved

→ **83%**
Adverse impact ratio

Bell Analytics

Agenda

- Introduction
- Relevant scope
- Quick thoughts on genAI
- Our framework
  - Performance monitoring
  - Unfair discrimination testing
- Range of real-world outcomes
- Q&A

Elaine Gibbs
CEO & Co-Founder, Bell Analytics

epg@bell-analytics.com | LinkedIn

Accessible, expert technical guidance on AI/ML strategy, adoption, and governance

- Unfair discrimination testing
- Technical “phone-a-friend”
- Thought leadership and education

Bell Analytics

Drivers of current testing activity

Jurisdictions

NAIC

Risk management guidelines

Evaluation tool

Risk management requirements

Test requirements [draft]*

Risk management requirements*

Test guidelines*

Technology

Predictive models

"ECDIS"*

Novel or non-traditional consumer data

* External consumer data and information source.

Use Case

Influence or drive consumer decision

With possibility of adverse consumer impact

Links: NAIC AI Principles; NAIC Model Bulletin; NAIC AI Evaluation Tool (4.0); Colorado Senate Bill 21-169 and related regulations; New York Insurance Circular Letter No. 1 (2019) and Insurance Circular Letter No. 7 (2024). More on individual state adoption of the NAIC Model Bulletin here: https://www.bell-analytics.com/blog/adoption-of-naic-model-bulletin * These requirements are specific to models and data used in pricing and underwriting only.

Bell Analytics

Scope in practice

Concern

Performance
Accuracy, drift

Protected class unfair discrimination
Race/ethnicity, gender

Technology

External score
e.g., credit-based insurance score, marketing response score

Internal score
including input variables

Use Case

Underwriting
Accelerated triage, risk class assignment

Claims
Fraud detection, lapse

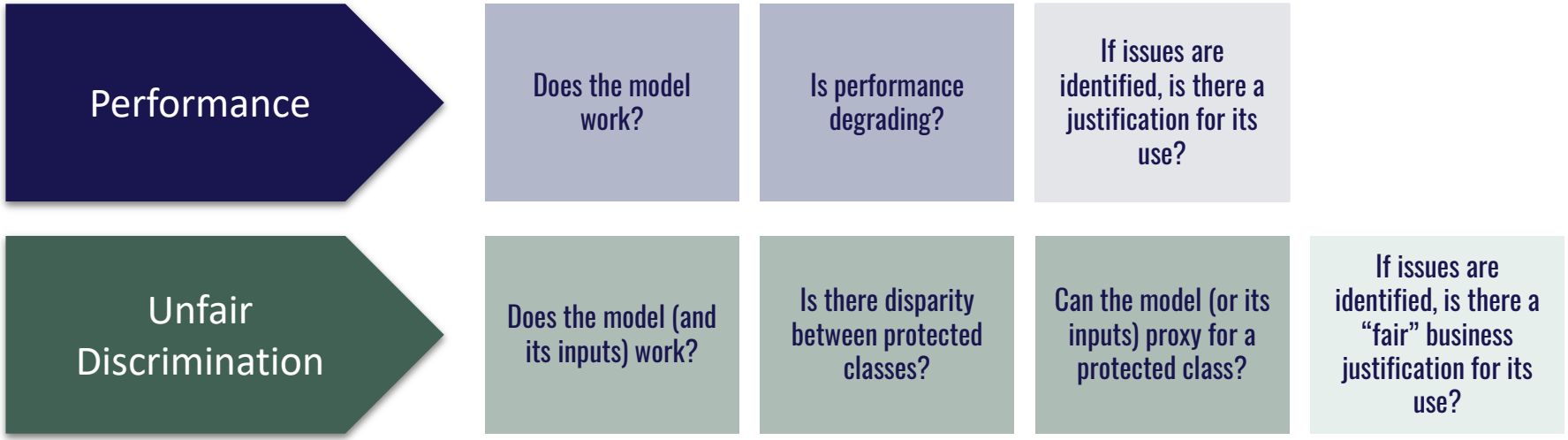
Marketing
New prospect, cross-sell

Quick thoughts on genAI

- Testing/monitoring **even more nascent** in terms of best practices
- May be **less of a focus** because applications (to date) have less impact on important consumer outcomes
- **Performance monitoring** often the priority
- Consider leveraging **existing human audit** processes

Bell Analytics

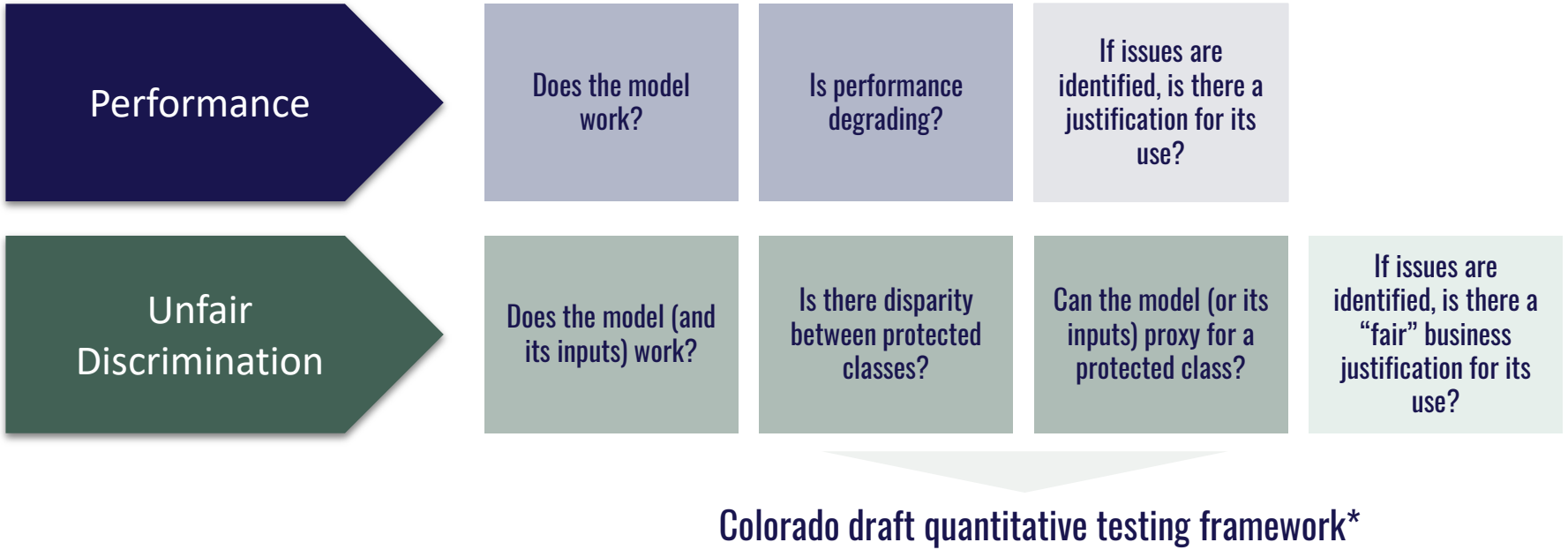
Our framework: The questions we ask (and answer)



** We say "model" here. This framework also applies to specific datapoints, if those are used on a standalone basis.

Bell Analytics

Our framework: The questions we ask (and answer)



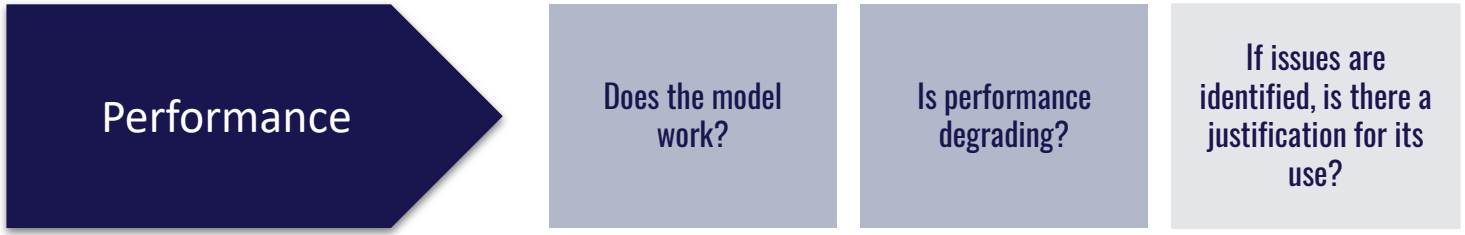
** We say "model" here. This framework also applies to specific datapoints, if those are used on a standalone basis. * If Colorado's draft quantitative testing framework passes, carriers using ECDIS to underwrite or price will have a prescribed testing approach. The draft regulation addresses these two questions in a highly specified manner. Otherwise, no prescribed framework or standards exist to determine the answer to any of these questions.

We use the terms “model,” “score,” and “risk score” interchangeably

Our insurer uses a **risk score** to triage certain claims for additional scrutiny

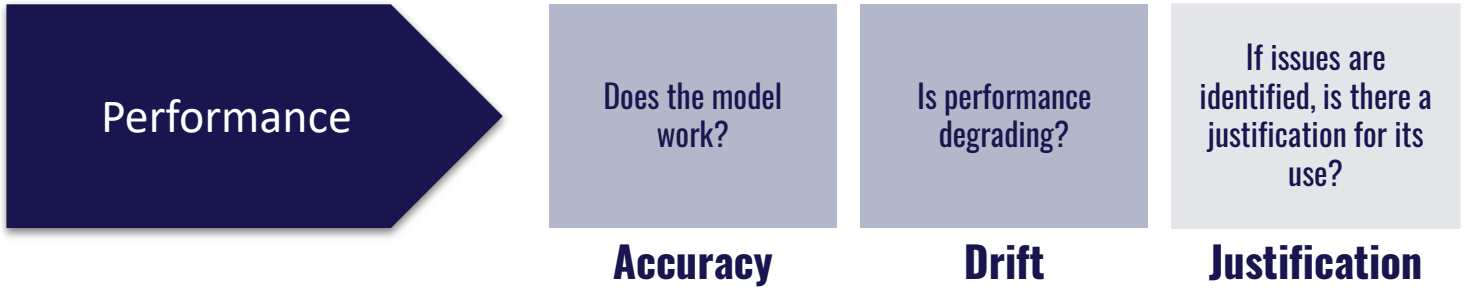
Bell Analytics

Let's start with performance monitoring

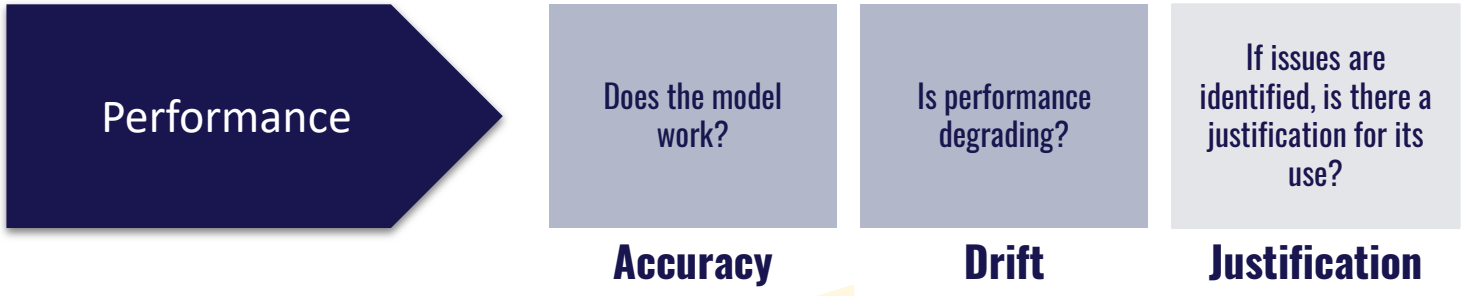


Bell Analytics

Let's start with performance monitoring



Let's start with performance monitoring



The business may be tracking these frequently, but only formally documenting them on a [quarterly / twice yearly / annual] basis

Accuracy: Does the model work?

Two primary approaches:

1. **Aggregate measures of error**

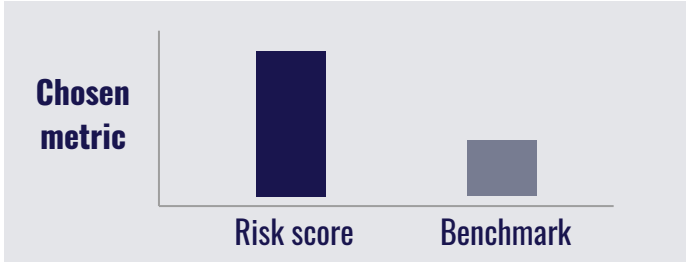
2. **Segmented measures of error ("Lift Chart")**

Bell Analytics

Accuracy: Does the model work?

Two primary approaches:

1. **Aggregate measures of error**



"75% of claims selected by the risk score truly needed the scrutiny, vs. 50% in the human-selected process"

Measures like: accuracy, precision, recall, area under the curve ("AUC"), or operational KPI's specific to the use case

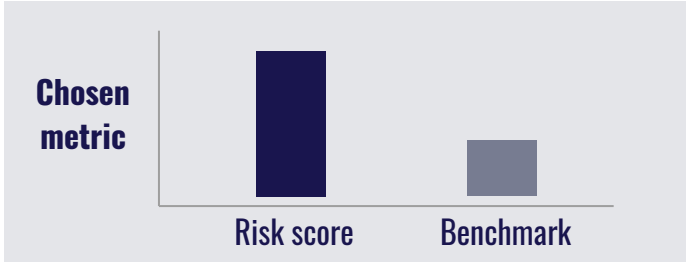
2. **Segmented measures of error ("Lift Chart")**

Bell Analytics

Accuracy: Does the model work?

Two primary approaches:

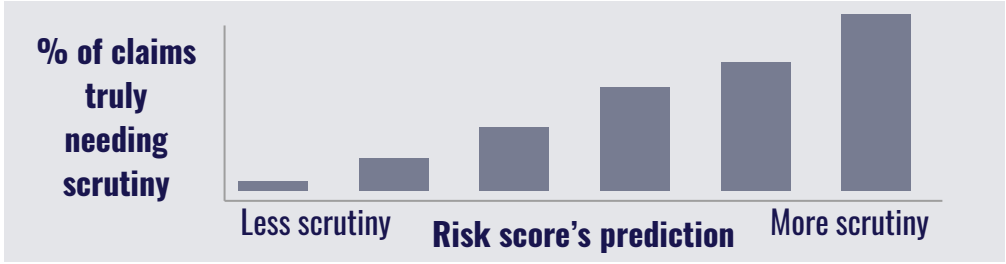
1. **Aggregate measures of error**



"75% of claims selected by the risk score truly needed the scrutiny, vs. 50% in the human-selected process"

Measures like: accuracy, precision, recall, area under the curve ("AUC"), or operational KPI's specific to the use case

2. **Segmented measures of error ("Lift Chart")**



Drift: Is performance degrading?

The answer is (often) YES

Note: Drift is defined in the NAIC Model Bulletin on AI as: Decay of a model's performance over time arising from underlying changes such as the definitions, distributions, and/or statistical properties between the data used to train the model and the data on which it is deployed.

Drift: Is performance degrading?

The answer is (often) YES

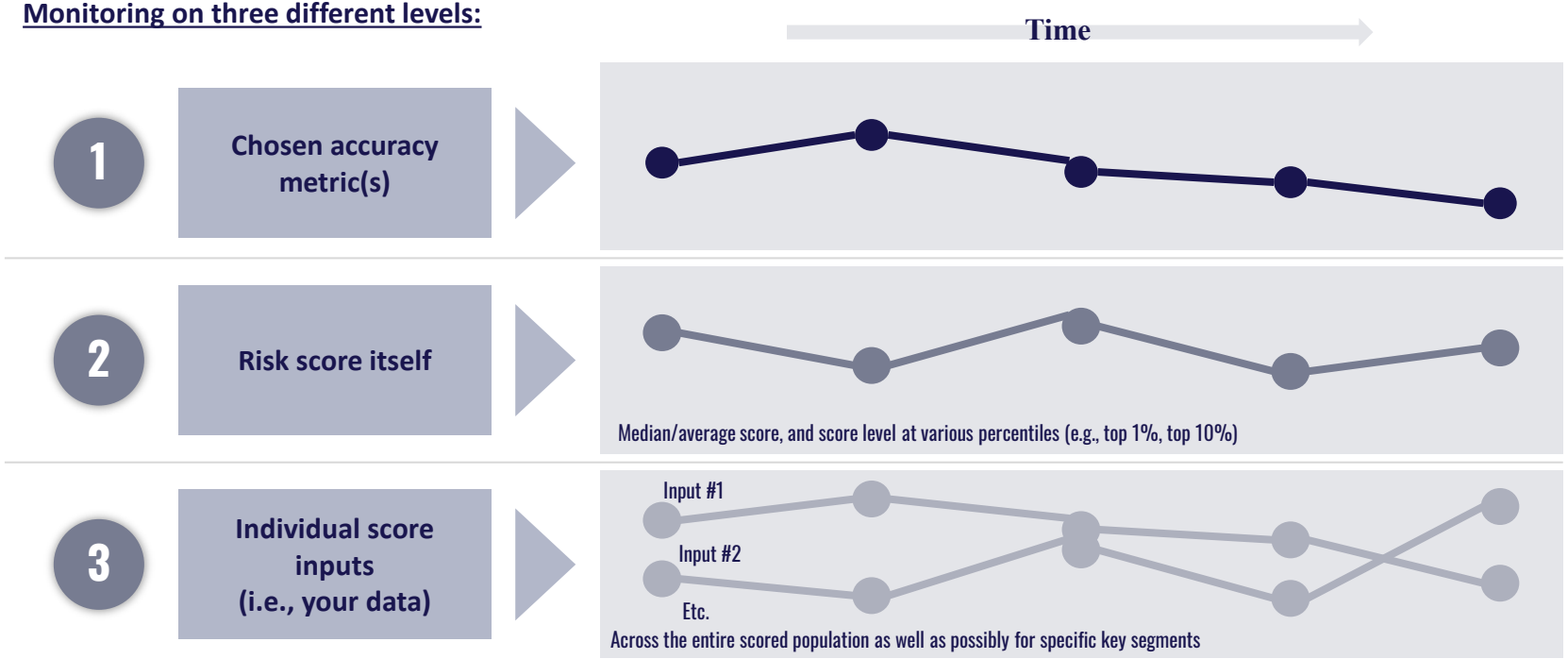
Short-term concern:
Your data is messed up & needs immediate fixes

Long-term concern:
The world has changed & the model is stale

Note: Drift is defined in the NAIC Model Bulletin on AI as: Decay of a model's performance over time arising from underlying changes such as the definitions, distributions, and/or statistical properties between the data used to train the model and the data on which it is deployed.

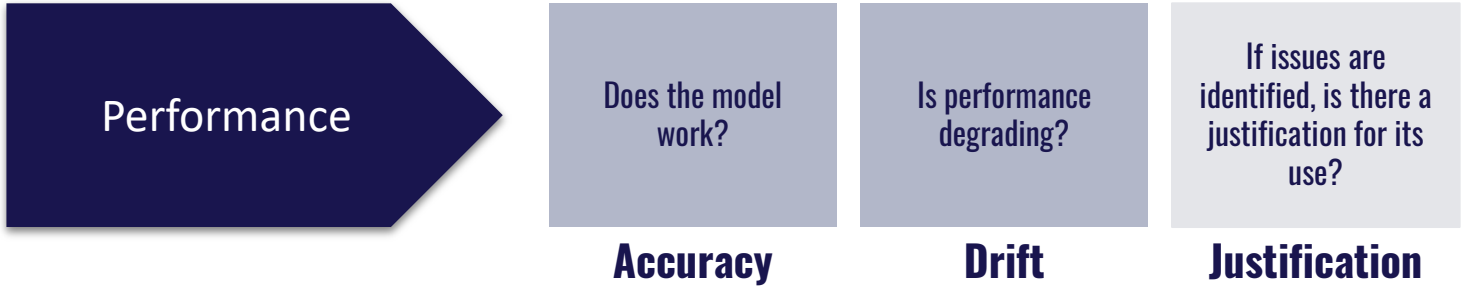
Drift: Is performance degrading?

Monitoring on three different levels:



Note: Drift is defined in the NAIC Model Bulletin on AI as: Decay of a model's performance over time arising from underlying changes such as the definitions, distributions, and/or statistical properties between the data used to train the model and the data on which it is deployed.

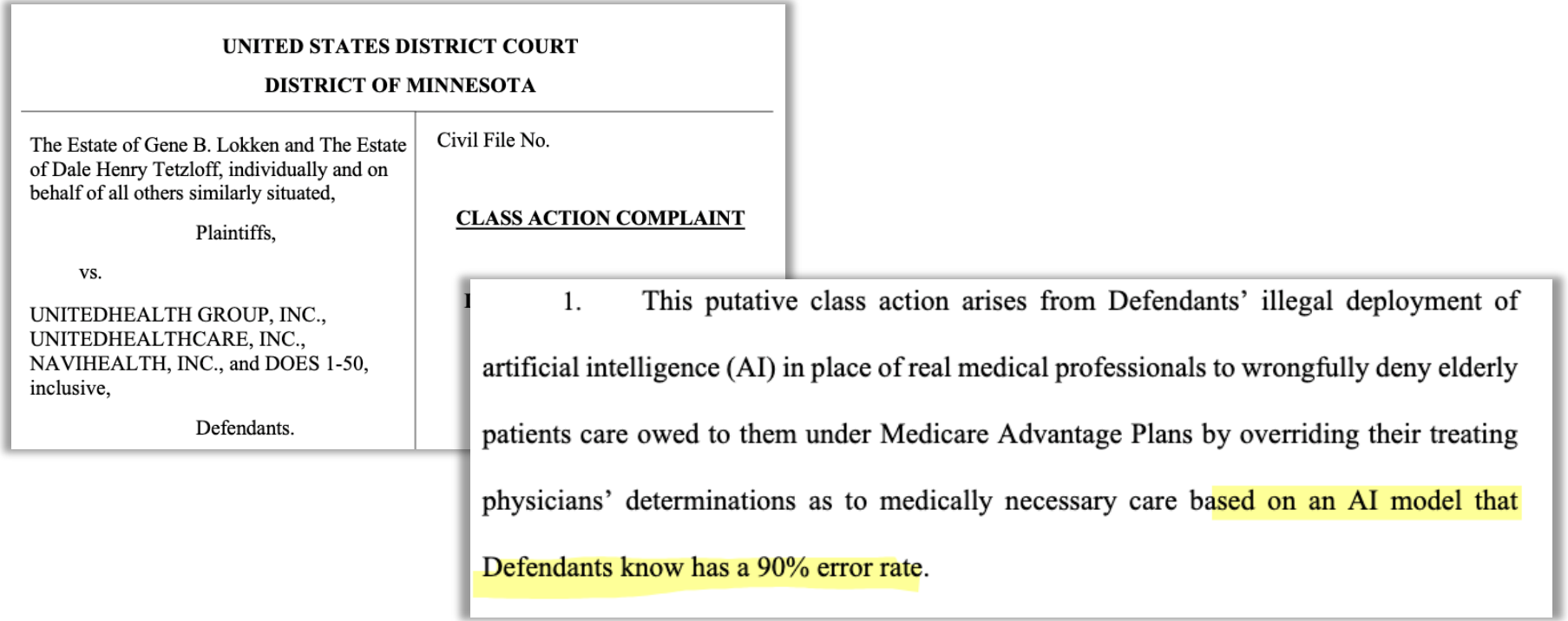
Justification: Interpret results



- ✓ Summarize findings
- ✓ Reference relevant benchmarks
- ✓ Explain aberrant results

Bell Analytics

Performance monitoring, benchmarks matter



UNITED STATES DISTRICT COURT

DISTRICT OF MINNESOTA

The Estate of Gene B. Lokken and The Estate of Dale Henry Tetzloff, individually and on behalf of all others similarly situated, Plaintiffs, vs. UNITEDHEALTH GROUP, INC., UNITEDHEALTHCARE, INC., NAVIHEALTH, INC., and DOES 1-50, inclusive, Defendants.	Civil File No. **CLASS ACTION COMPLAINT**

1. This putative class action arises from Defendants' illegal deployment of artificial intelligence (AI) in place of real medical professionals to wrongfully deny elderly patients care owed to them under Medicare Advantage Plans by overriding their treating physicians' determinations as to medically necessary care based on an AI model that Defendants know has a 90% error rate.

Note: 90% error rate quoted by plaintiff's lawyers was based on the percentage of claims that went to appeal that were ultimately overturned. Original complaint source.

Bell Analytics

Onto unfair discrimination

Performance	Does the model work?	Is performance degrading?	If issues are identified, is there a justification for its use?	
Unfair Discrimination	Does the model (and its inputs) work?	Is there disparity between protected classes?	Can the model (or its inputs) proxy for a protected class?	If issues are identified, is there a “fair” business justification for its use?

Bell Analytics

Onto unfair discrimination

This analysis is typically done on a [twice yearly / annual] basis

Bell Analytics

Step #1: Race and ethnic class inference

Name + location inference

- Standard algorithms (i.e., BISG and BIFSG) use first and/or last name and zip code as inputs
- Output is, for every individual, the likelihood of belonging to a set of race & ethnic classes

Four testable race and ethnic classes

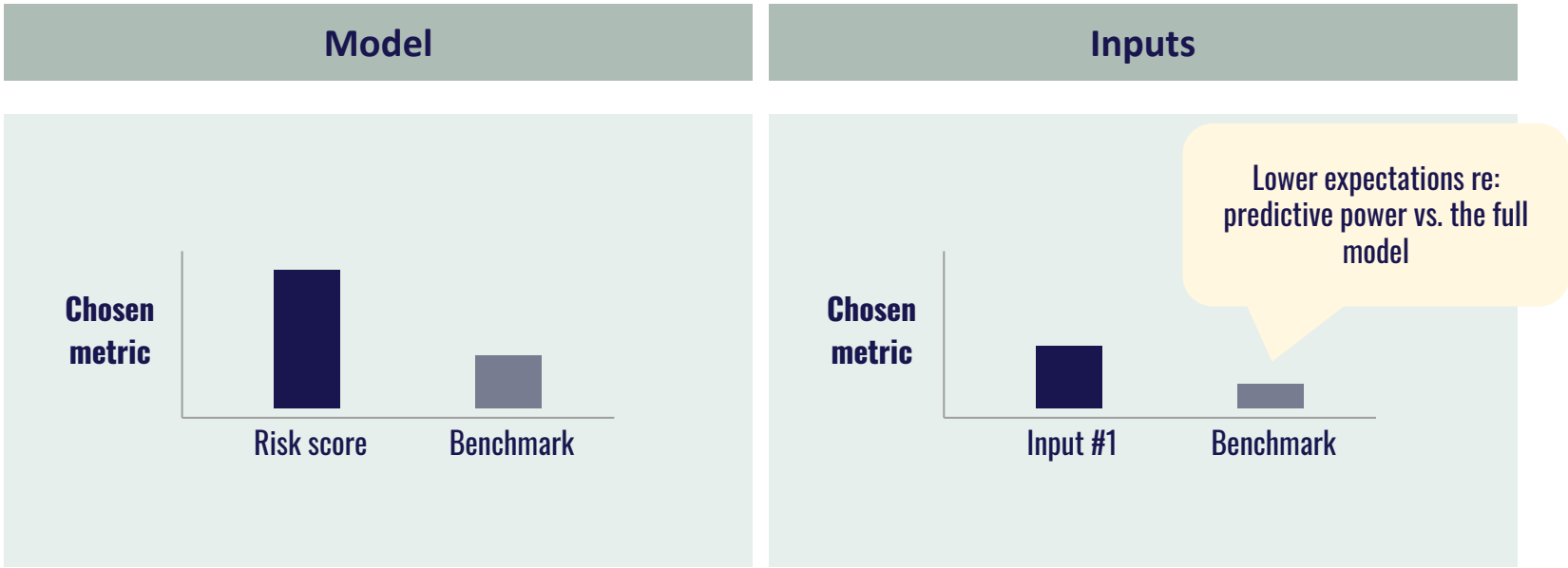
- (1) White; (2) Black; (3) Hispanic; (4) Asian or Pacific Islander
- Tend to exclude Native and Multi-Racial given accuracy and sample-size concerns
- Disparities calculated vs. the majority class (often White) as reference

Inference choices matter

- Range of approaches for how to incorporate probabilities
- Choices impact population coverage as well as accuracy, which skew analyses in various directions. Implement thoughtfully with use case in mind

Resources: Summary and deep-dive research from Bell Analytics.

Accuracy: Does the model (and its inputs) work?



Disparity: Disparity between protected classes?

A range of approaches:

1. **Simple Adverse Impact Ratio**

80% of dog people pass without extra scrutiny

70% of cat people pass without extra scrutiny

Cat people pass at **87.5%** the rate of dog people*

2. **Adjusted Adverse Impact Ratio**

3. **Modeled Disparity**

* 70% / 80% = 87.5%.

Bell Analytics

Disparity: Disparity between protected classes?

A range of approaches:

1 **Simple Adverse Impact Ratio**

80% of dog people pass without extra scrutiny

70% of cat people pass without extra scrutiny

Cat people pass at **87.5%** the rate of dog people*

2 **Adjusted Adverse Impact Ratio**

Dog & cat people vary **by age**

78% of cat people pass, on equal age footing

Cat people, age-adjusted, pass at **97.5%** the rate of dog people**

3 **Modeled Disparity**

* 70% / 80% = 87.5%. ** 78% / 80% = 97.5%.

Bell Analytics

Disparity: Disparity between protected classes?

A range of approaches:

1 **Simple Adverse Impact Ratio**

80% of dog people pass without extra scrutiny

70% of cat people pass without extra scrutiny

Cat people pass at **87.5%** the rate of dog people*

2 **Adjusted Adverse Impact Ratio**

Dog & cat people vary **by age**

78% of cat people pass, on equal age footing

Cat people, age-adjusted, pass at **97.5%** the rate of dog people**

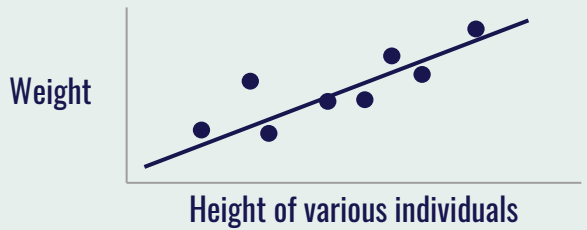
3 **Modeled Disparity**

*Claim Scrutiny Decision ~ AGE + POLICY SIZE + CAT PERSON (Y/N)****

* 70% / 80% = 87.5%. ** 78% / 80% = 97.5%. *** This approach fits an equation to past decisions, quantifying the relationship between those decisions and the selected factors, including our protected class. In this example, the influence ascribed to CAT PERSON (Y/N) indicates the disparity associated with membership in that class, after controlling for age and policy size.

Bell Analytics

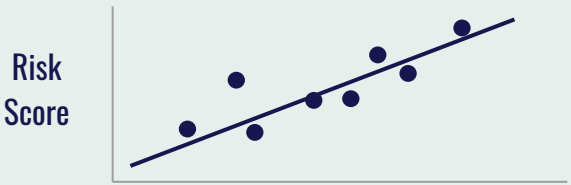
Proxy: Can model (or its inputs) stand-in for protected class?

Simple	More Involved
Correlation metric (r):* 	

* This is the correlation metric (r) you may remember from high school math class. It's a measurement of how closely two variables are related, ranging between -1 and 1. 1 means perfectly in sync. 0 means no relationship whatsoever. -1 means perfectly in sync, but in opposition (one goes up, the other goes down). Things get a little more complicated with categorical variables, but equivalent metrics (like Cramer's V) can be leveraged there in a similar fashion.

Bell Analytics

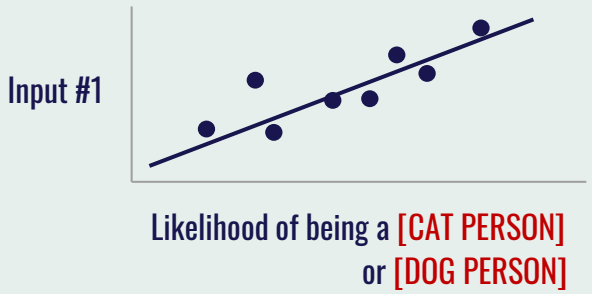
Proxy: Can model (or its inputs) stand-in for protected class?

Simple	More Involved
Correlation metric (r):*  Likelihood of being a [CAT PERSON] or [DOG PERSON]	

* This is the correlation metric (r) you may remember from high school math class. It's a measurement of how closely two variables are related, ranging between -1 and 1. 1 means perfectly in sync. 0 means no relationship whatsoever. -1 means perfectly in sync, but in opposition (one goes up, the other goes down). Things get a little more complicated with categorical variables, but equivalent metrics (like Cramer's V) can be leveraged there in a similar fashion.

Bell Analytics

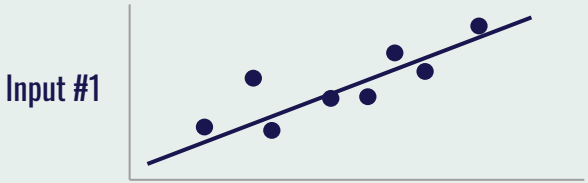
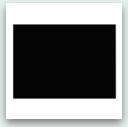
Proxy: Can model (or its inputs) stand-in for protected class?

Simple	More Involved
Correlation metric (r):* 	

* This is the correlation metric (r) you may remember from high school math class. It's a measurement of how closely two variables are related, ranging between -1 and 1. 1 means perfectly in sync. 0 means no relationship whatsoever. -1 means perfectly in sync, but in opposition (one goes up, the other goes down). Things get a little more complicated with categorical variables, but equivalent metrics (like Cramer's V) can be leveraged there in a similar fashion.

Bell Analytics

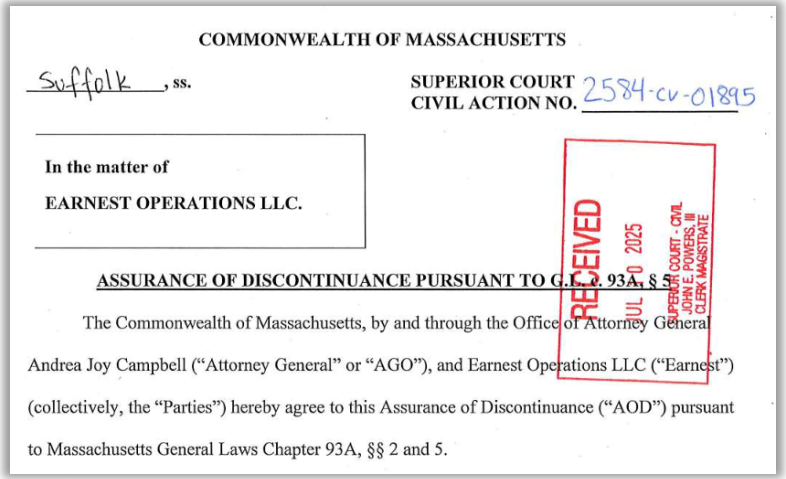
Proxy: Can model (or its inputs) stand-in for protected class?

Simple	More Involved
Correlation metric (r):*	Modeled approach:**
	
Likelihood of being a [CAT PERSON] or [DOG PERSON]	How much of the Risk Score's influence on our consumer decision is based on protected class membership, after controlling for other factors?

* This is the correlation metric (r) you may remember from high school math class. It's a measurement of how closely two variables are related, ranging between -1 and 1. 1 means perfectly in sync. 0 means no relationship whatsoever. -1 means perfectly in sync, but in opposition (one goes up, the other goes down). Things get a little more complicated with categorical variables, but equivalent metrics (like Cramer's V) can be leveraged there in a similar fashion. ** More detail on Colorado's draft quantitative testing protocol here (original) and here (ACLI version).

Bell Analytics

Proxies in action



COMMONWEALTH OF MASSACHUSETTS

Suffolk, ss.

SUPERIOR COURT
CIVIL ACTION NO. 2584-cv-01895

In the matter of

EARNEST OPERATIONS LLC.

RECEIVED
JUL 10 2025
SUPERIOR COURT - CIVIL
JOHN E. POWERS, III
CLERK MAGISTRATE

ASSURANCE OF DISCONTINUANCE PURSUANT TO G.L. c. 93A, § 5

The Commonwealth of Massachusetts, by and through the Office of Attorney General Andrea Joy Campbell ("Attorney General" or "AGO"), and Earnest Operations LLC ("Earnest") (collectively, the "Parties") hereby agree to this Assurance of Discontinuance ("AOD") pursuant to Massachusetts General Laws Chapter 93A, §§ 2 and 5.

38. The SLR Model included an evaluation of the applicant's Cohort Default Rate ("CDR") as a Weighted Input until September 13, 2017.

39. The Cohort Default Rate is produced by the U.S. Department of Education and describes the average rate of loan defaults associated with specific higher education institutions.

41. Earnest's use of the CDR subscore in its SLR Underwriting Model resulted in disparate impact in approval rates and loan terms in the SLR product, with Black and Hispanic applicants more likely to be penalized than White applicants.

42. The Attorney General alleges that each time Earnest used the CDR variable in Underwriting, it violated ECOA, and thereby violated G.L. c. 93A, § 2.

Source: Settlement. Note: Example above is in the consumer credit / lending business. Certain legal standards, like disparate impact and ECOA, are not directly translatable to insurance. Presented illustratively only to show where possible proxy inputs can come into play.

Justification: If issues flagged, is there a "fair" business reason?

Common fact patterns:

Some level of disparity

- Disparity disappears after accounting for other factors
- Model nor inputs appear to act as proxies
- Model cannot be replaced by traditional factors
- Model works well for each individual class

Possible proxy

- Clear relationship to relevant risk (i.e., high accuracy)
- No harm to related class from use of input or model
- Input functionally unable to act as proxy

Whether remediation is required is a judgment call based on the strength of these fact patterns

Remediation could include: 1) limit model scope; 2) remove troubling inputs; 3) rebuild model; 4) remove model

Real-world testing findings

✅ Everything looks great!

✅ Disparity found between race/ethnic groups, but differences are fully explained by well-established risk factors

🤔 Disparity found between race/ethnic groups, with differences partially explained by well-established risk factors. Risk score highly predictive and works for individual groups

🤔 Risk score input moderately proxies for a race/ethnic group, but in a way that does not drive better or worse outcomes

❌ Risk score input strongly proxies for race/ethnic groups, and does not have much predictive power at all

❌ Risk score actually doesn't work very well!

Q & A

Thank You!

Thank you, Elaine!

Administrative Reminders

- Please complete the 1-minute post event survey you will receive.

- Post Event Communication
  - The presentation deck
  - A link to the recording
  - A “Certificate of Attendance” template (for those who attended the live event and who wish to self-submit for CLE or CE) with the organizations they are involved with

CLE Information

While CEFLI does not file its materials with State Bar Associations, if you plan to self-submit for potential CLE consideration, the following may be helpful:

- CEFLI is the sponsor of this event.
- Only those who attend the live webinar receive the Certificate of Attendance form.
- CEFLI does not have a way to know how many attorneys attended the event.
- CEFLI's 1-hour webinars do not have a timed agenda.
- Participants have the ability to ask questions during CEFLI webinar events.
- CEFLI is not a marketing organization. It is a compliance and ethics organization whose mission is to support professionals by providing educational opportunities that address current compliance matters.

Bell Analytics

Data Science Done Right

Elaine Gibbs
epg@bell-analytics.com