Better Government Lab

McCourt School of Public Policy GEORGETOWN UNIVERSITY | Ford School of Public Policy UNIVERSITY OF MICHIGAN

QC Data Workshop: Predictive Analytics for SNAP PER Reduction

March 2, 2026 | Workshop Summary

Overview

On March 2, 2026 The Better Government Lab at Georgetown and University of Michigan, SCALE Lab at Yale, and Aspen Financial Security Program convened a group of nearly 40 research and data analytics staff from across 15 states involved in SNAP QC data modeling.

The purpose of the gathering was to foster collaboration and share modeling practices among state agencies in response to HR1's new SNAP PER cost share requirements. The conversation was guided by researchers from Georgetown and Yale sharing examples from their work, with active discussion from participants around key questions:

- What specific tools do you use most often?
- Which outcomes do you model?
- What variables have you found helpful?
- What strategies have you used to improve precision for flagging cases?
- How do you assess models for fairness?

Participant experience ranged from those in the planning phase of modeling to others modeling active case data in collaboration with caseworkers and QA staff for review and correction. This document summarizes key takeaways from this discussion.

SCALE

FINANCIAL SECURITY PROGRAM aspen institute

Better Government Lab

GU McCourt School of Public Policy GEORGETOWN UNIVERSITY | Ford School of Public Policy UNIVERSITY OF MICHIGAN

What specific tools do you use most often?

Category	Tool
Coding Languages & Analysis Tools	• R • Python (e.g., with scikit-learn) • Stata
Statistical Methods & Machine Learning Algorithms	• Random Forest (RF) • Boosted Trees/Bagging (BG) • Logistic Regression (Logit) • XGBoost
External Programs/Platforms	• Oracle APEX • Gemini • Slides

Example:
In one state, a risk model has been developed to assist the Quality Assurance (QA) team and state SNAP workers in identifying cases for further review. The model analyzes discrete indicators - such as mortgage or principal payments, health insurance premiums - that end in round numbers. Focusing on discrete patterns has been helpful for case identification, while cases with non-discrete patterns are routed to the QA team for review. The model uses XGBoost and is designed to maximize precision for class 1 cases, so the QA team receives high-confidence cases that are likely to benefit from review.

The model is not embedded in the eligibility system, instead it's accessed through a third party interface, Oracle Apex, where caseworkers can find a case using a client ID, and receive information about which factors have been flagged for review. The system runs on R and Python and requires a weekly manual refresh. Reviewing flagged cases is a request rather than a requirement for caseworkers, highlighting that the effectiveness of these tools will depend on how they are implemented.

Which outcomes do you model?

Before starting with modeling it's important to think through which outcome matters most, whether it makes sense to focus on a subset of errors, and what the denominator should be. If the priority is to reduce the payment error rate (PER) then most models should be intentional about how errors equal to or below the error tolerance threshold ($58 for FY2026) are treated.

Participants spoke at length about how the error threshold and review process influences their approach to modeling. A brief overview of the PER and error threshold review process is as follows:

- **Comparison 1**: QC checks whether the allotment issued differs from what should have been issued based on verification of sample month data. If the difference exceeds $58, the review moves to Comparison 2.
- **Comparison 2**: QC reviews circumstances at certification. If this comparison also shows a difference greater than $58, the case is reported as a payment error and included in the PER.
- In most cases the lower of Comparison 1 and Comparison 2 will be recorded.

Participants shared focusing on the following outcomes:

- **Errors greater than the threshold:** Most participants focus on errors greater than the threshold, noting that because of the structure of the review process, smaller errors rarely receive additional scrutiny and may create misleading signals.
- **Presence of any error including cases under threshold:** Some workshop participants are modeling all errors, noting that smaller errors can be helpful to find trends, may add value to classification-type models, and help maintain the signal between big errors and no errors.
- **Actionable errors on active cases:** Some noted focusing on errors that can be addressed in real time using caseworker validation and simple binary error signals for lighter-weight models.
- **Magnitude and timing:** Workshop participants noted focusing on magnitude of errors as an outcome, as well as using QC data to identify when errors occurred.

Modeling choices

- **Predicting agency and client errors separately**: Given that audit, review, and business processes changes will be different for agency versus client errors, these are often modeled separately.
- **Emphasis on precision over recall:** given time constraints, most participants were focused on precision over a model that captures everything. Some participants looked at the challenge as optimizing error dollars prevented per hour of review time.

Augmenting QC data

Some participants reported that bringing in richer data yielded more predictive power and insights about potential process and system change:

- **Incorporating insights from caseworker reviews** building models with insights from staff about how errors had occurred helped improve predictive power.
- **Adding data on process, especially for agency errors,** such as variables to capture the complexity of a case and the tasks required to process it from the worker perspective. For example, adding task management data into models to identify what might lead a caseworker to fail to apply a change or to look for verification documents.
- **Adding other data such as case notes, RFIs**

What variables have you found helpful?

Federal QC data are small (~1,000 cases per year) so it's beneficial to encode institutional knowledge through curated variable selection. QC only has a few thousand cases per year, leading to a data starved environment for predictive modeling. What had worked for agencies:

- **Creating feature derivatives**, such as the number of income types, the presence of self-employment income, or the number of deduction types, flagging that certain amounts are round numbers, which has been more predictive than using raw data alone.
- **Variables that most improved model performance:**
    - Shelter costs exceeding 50% of income
    - Benefit amount relative to max allotment (ratio 0–1; seems to help with client-caused)
    - Number of sources of income, number of types of deductions
    - Months since last certification
    - Indicators combining homelessness status with shelter deduction
- **Data and variables to spot potential agency errors**:
    - Number of different screens the caseworker had to hit as an agency error predictor
    - How many income types, non-direct data sources that could influence errors
    - Worker experience:
        - One state found that worker experience was a top variable. More experience -> less likely to have an error. People talked about ways that segmenting this by type of error might be helpful, such as those where complexity and experience might predict positive effects versus those where more recent training might be beneficial. This could assist with targeted retraining as well. Participants noted to be mindful of worker protections and non-punitive enforcement.
        - Another state did not find that experience was predictive - they could see that junior caseworkers were allocated simpler tasks, but apart from that they didn't find anything significant.

- **Number of months since recertification / number of days until next recertification:** This was identified for states that use simplified interim reporting.

What strategies have you used to improve precision for flagging cases?

In addition to selecting the right outcomes, cases, and constructing new variables for each model, participants shared several strategies for improving model precision:

Start by focusing on highest-impact errors

- Participants noted income as the biggest driver of errors, suggesting the review of cases with 3 or more household members, and over $1,500 in income. Larger errors are easier to identify for larger households – a $58 error would be a much bigger miss for a household of 1 or 2. Increased household size also implies potentially higher case complexity. One state identified secondary review at case wrap-up for households over 3 members.
- The second largest driver of errors identified by the group were errors in eligibility (eligible vs ineligible), which is often client-caused due to incomplete or inaccurate information disclosed during their interview.

Refine variables

- If expense to income ratio is less effective as a summary variable, consider whether it adds value to the model.
- Depending on the type of model, remove redundant variables, such as measures of income at various levels of aggregation.
- Indicators that combine homelessness status and shelter deduction have shown high precision.
- For client-caused errors, benefit ratio seems to be an effective flag.

Integrate secondary human review

- Several participants noted that secondary reviewers almost always catch errors and recommended enlisting seasoned leads who have experience auditing.
- Given time constraints, secondary reviewers should not be tasked with reviewing an entire case. Participants recommended providing context on the likely error source and focusing reviewers on a few key data points — one participant noted that reviewing income alone took the average reviewer about 5 minutes. Other variables that could be isolated for human review include shelter deduction, household composition, able-bodied designation, and citizenship.
- Participants identified a goal of fixing the most errors or error dollars per unit of time, noting that data modelers should be aware of the trade-off between more intensive reviews and time costs.

- Ideally, cases would be run through a model before certification or recertification finalization to identify high-risk cases that could be sent to a supervisor for review. Those that weren't flagged would go straight through to processing.
- One state implemented secondary review at case wrap-up for households over 3 members.

Considerations for AI-assisted review

- AI tools can find discrepancies between interview notes and what is documented in the casefile, flagging cases where verifications are incomplete or information is misaligned.
- However, many errors are revealed from information that is *not* in the case file. Modelers should be aware that interviewers may be reluctant to ask some questions during interviews and that AI tools may miss these gaps entirely.

How do you assess models for fairness?

As models become increasingly complex, there can be a risk that they disproportionately allocate scrutiny across groups. While this topic was top of mind for participants, most identified being early in their data modeling journey – eager to learn from their peers, but not yet at the stage of actively correcting for bias.

Participants shared early thoughts on assessing models for fairness including:

- Ensure the model is calibrated and that "risk" means the same thing for each group.
- Assess fairness for the metric(s) that matter – e.g., False Positive Rate
- Focus on high precision approaches that stay as close to the baseline rate as possible, reducing the risk of flagging large numbers of households for additional scrutiny.
- Avoid imposing additional burdens on benefit recipients, such as additional verifications or secondary interviews.
- Federal QC data is close to a random sample, reducing concerns of fairness when modeling. When building your own model, be thoughtful about the selection process of the data that's feeding your system given that data elements can be biased and encode existing bias.
- One state noted that within the same household size, they did not see bias.

Additional Discussion

Partnerships and data sharing

- Some participants have had success raising private funds for doing this work.
- Participants noted that partnerships with quality improvement teams have been successful.
- For those just beginning data modeling, they see a golden opportunity to work closely with QA teams and develop strong feedback loops.
- For those working with external partners, participants encouraged the use of data sharing agreements (DSA) or data use agreements (DUA), and embedding contractors within the state

to work in a protected and credentialed environment (no sharing of PII outside of state machines).

Barriers and challenges

- Many participants are just reaching the modeling phase.
- County administered programs create more fragmented data compared to state administered programs, increasing complexity with modeling and deployment.
- Some states are missing information about when errors occurred (no time period associated with errors) and see variations in data on who caused the error.
- The capacity of QA teams is often limited. Data teams want to ensure that they identify errors that are actionable and are not overly time consuming. Participants are grappling with whether a 5–minute QA review can actually result in finding and fixing errors and how to predict and measure the effectiveness of those interventions.

Looking ahead

- Should we be building separate models for certification vs recertification - agency vs client error, element or program factor?
- How do we ensure errors flagged are actionable?
- Assuming you could build the perfect model, what would you do with it and how?