"Improving Operational Agility at the IRS by Leveraging Intermediate Examination Results in a Sequential Decision-Making ML Pipeline"

Intermediate Outcomes

JSM

Brandon Anderson[1], Austin Miller[1], Alex Turk[1], Peter Henderson[2], Annette Portz[1]

August 8, 2024

[1] Internal Revenue Service, [2] Princeton University

Background

Key Concepts

Risk-Based Exam Selection

- Audit selection priority based on prediction models of tax non-compliance.

Feedback Loop

- Results of risk-based audits are used to train the models that select them.

Time-to-Insight

- The delay between case selection and outcome availability.

Concept Drift

- Real changes in tax policy and behavior that take place over time.

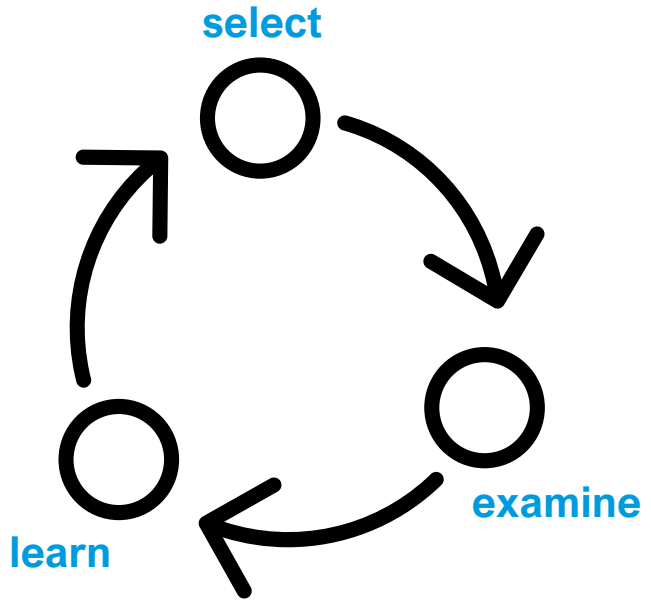
Intermediate Outcomes

- Case insights that can be made available before closure.

Approach

Overview

Ideally, **risk-based exam selection** pipelines exist as **feedback loops**, where insights from outcomes are used to improve risk models.

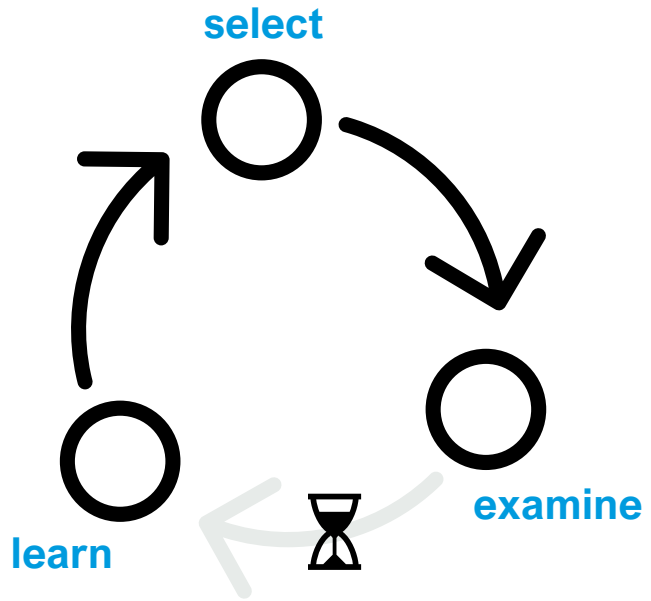


Approach

Overview

Ideally, **risk-based exam selection** pipelines exist as **feedback loops**, where insights from outcomes are used to improve risk models.

The examination process has a long **time-to-insight.** While we wait, the model is still making selections regardless of **concept drift**.



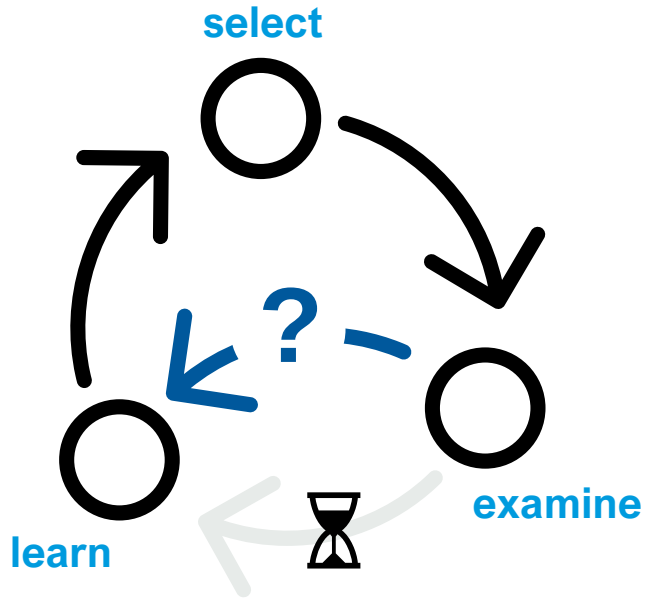
Approach

Overview

Ideally, **risk-based exam selection** pipelines exist as **feedback loops**, where insights from outcomes are used to improve risk models.

The examination process has a long **time-to-insight.** While we wait, the model is still making selections regardless of **concept drift**.

We don’t need to wait for closure to get data! Classification, grade-hours worked, related transactions, etc. – all relate to the eventual outcome. How can we use these **intermediate outcomes** help to bridge the gap?



Approach

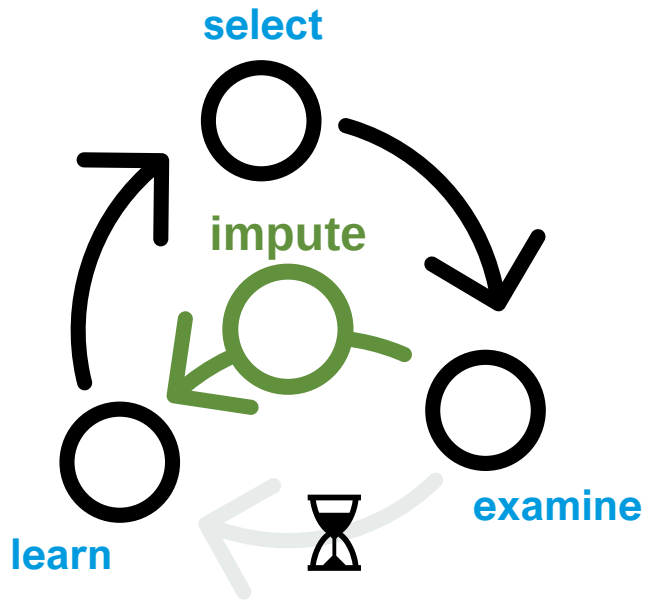
Overview

Ideally, **risk-based exam selection** pipelines exist as **feedback loops**, where insights from outcomes are used to improve risk models.

The examination process has a long **time-to-insight.** While we wait, the model is still making selections regardless of **concept drift**.

We don't need to wait for closure to get data! Classification, grade-hours worked, related transactions, etc. – all relate to the eventual outcome. How can we use these **intermediate outcomes** help to bridge the gap?

A. If we can train a more accurate model to predict adjustments using intermediate outcomes, we can **impute** training data for our selection model.



Approach

Overview

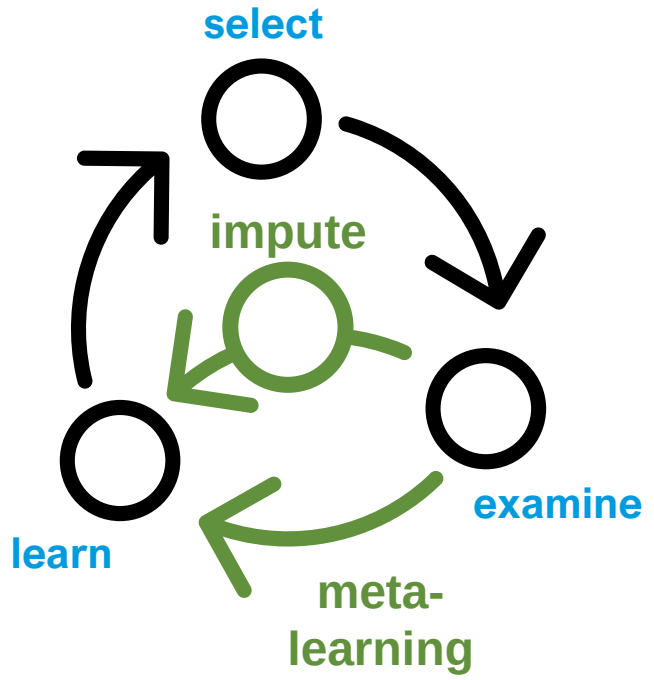
Ideally, **risk-based exam selection** pipelines exist as **feedback loops**, where insights from outcomes are used to improve risk models.

The examination process has a long **time-to-insight.** While we wait, the model is still making selections regardless of **concept drift**.

We don’t need to wait for closure to get data! Classification, grade-hours worked, related transactions, etc. – all relate to the eventual outcome. How can we use these **intermediate outcomes** help to bridge the gap?

A. If we can train a more accurate model to predict adjustments using intermediate outcomes, we can **impute** training data for our selection model.

B. Alternatively, we can train a model to predict outcomes at all stages of an exam, allowing us to perform continuous **meta-learning**.



Drilling Down

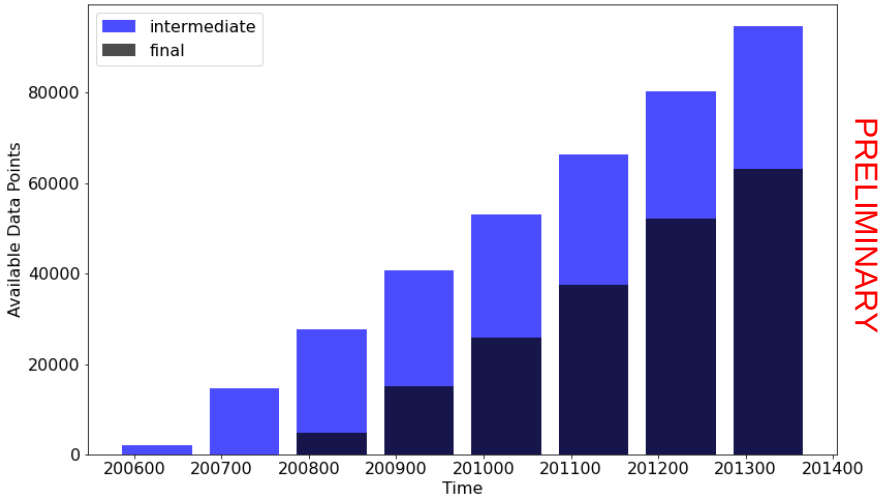
Intermediate Audit Outcomes

What are they?

- Case Classification – an expert looks at the case and flags how (and if) an audit should be pursued. Very early result.
- Exam Grade – the pay grade of the agent/officer assigned to the case. Very early result.
- Hours Worked – the number of hours worked on the case at any point. Available continuously.
- Days Open – the span of time that the case has been open. Available continuously.
- Case Transactions – transactions made by or to the auditee related to the tax module in question. Available sporadically (easiest to view cumulatively).
- Others . . .

When can we get them?

- Several intermediate results are available almost instantly.
- Contrast with full exam results that can take years to complete.



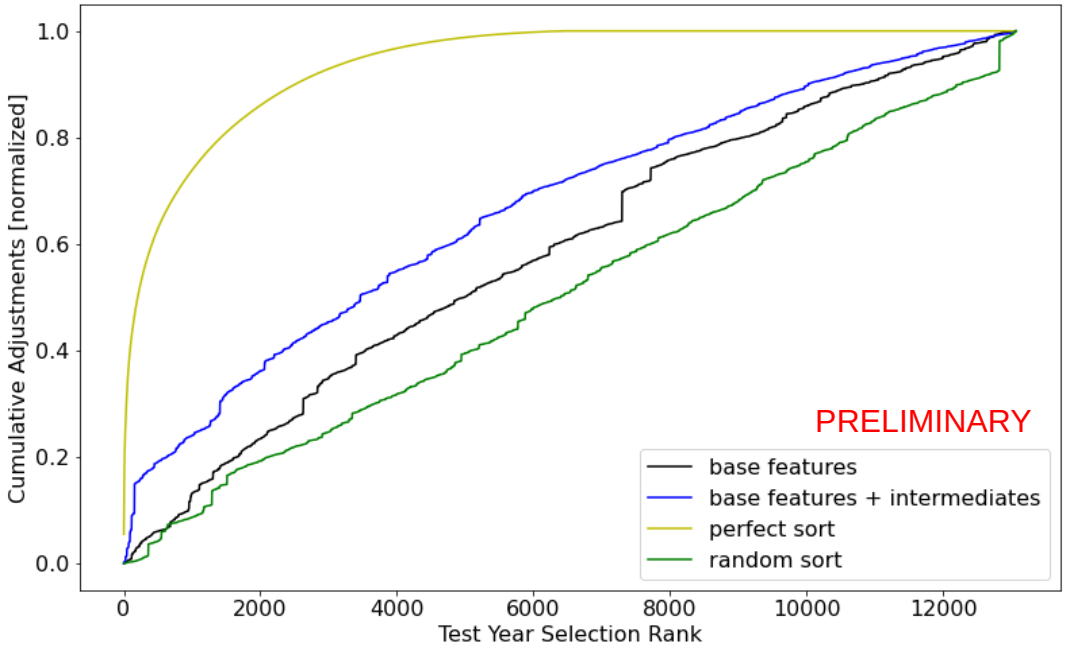
- Current sample designs wait until ALL cases are finished before updating the model. This increases the lag to several years.

Drilling Down

Intermediate Audit Outcomes

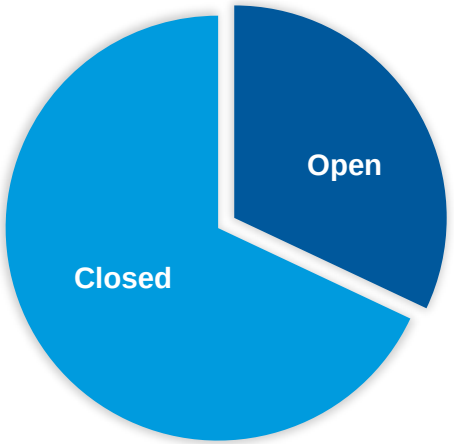
Are they predictive?

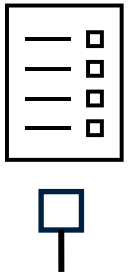
- Yes! Training a model with a feature set augmented by **intermediate outcomes does a better job at risk-sorting the test population**.
- This means we can make a better guess at how a case will end (unsurprisingly, accuracy improves as the case gets closer to completion).
- This might be useful by itself for strategic resource deployment, but as **none of these features are available at selection time, how can we make use of this to do better there?**



Drilling Down

Incorporating Insights





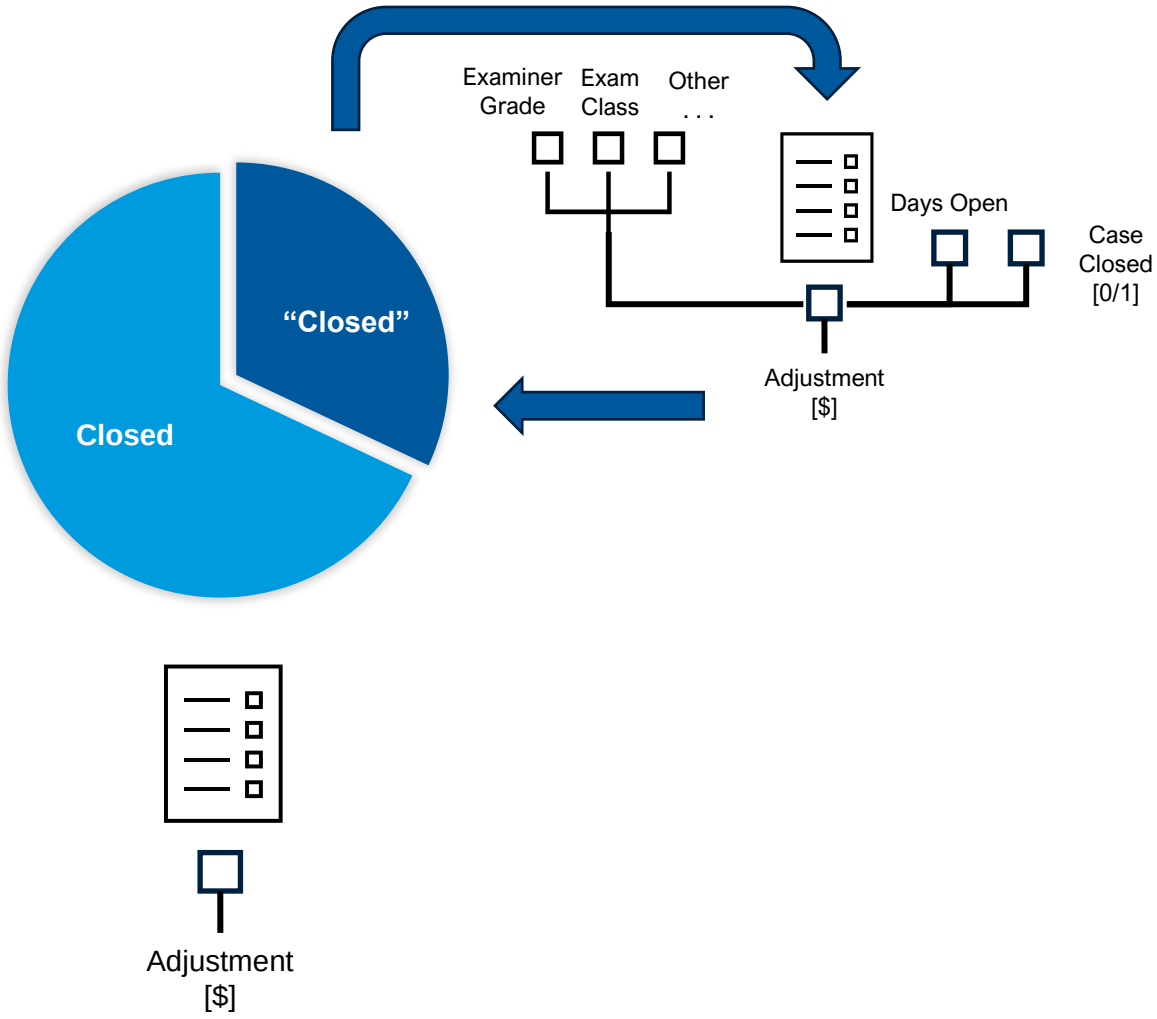
Adjustment
[$]

Drilling Down

Incorporating Insights

Imputation

- If the augmented model can predict outcomes that are closer to reality than our base model, we might move forward by synthesizing labels for open cases and incorporating them into our training data.
- Imputations should probably be done using some sampled uncertainty to avoid collapsing the true variance of cases.
- Some risk of overfitting.



Drilling Down

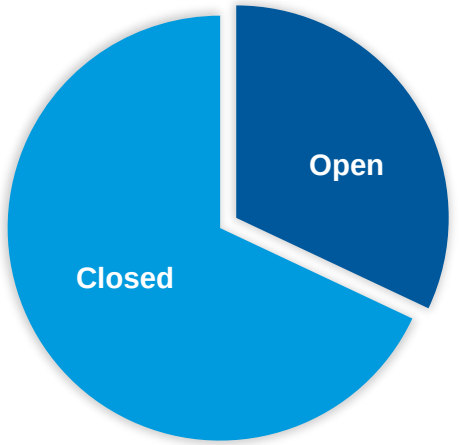
Incorporating Insights

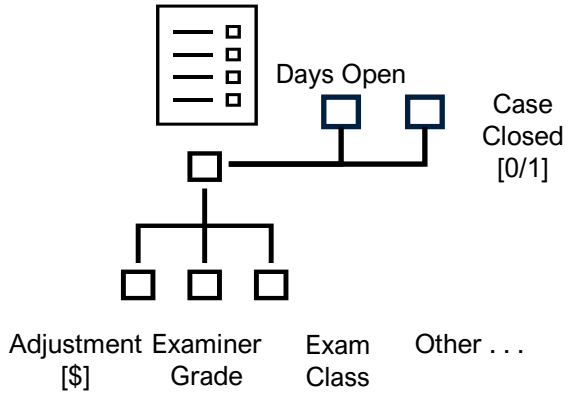
Imputation

- If the augmented model can predict outcomes that are closer to reality than our base model, we might move forward by synthesizing labels for open cases and incorporating them into our training data.
- Imputations should probably be done using some sampled uncertainty to avoid collapsing the true variance of cases.
- Some risk of overfitting.

Meta-Learning

- We can instead train a multi-target model to predict all outcomes of a case (including intermediates), given the time elapsed and closure status.
- Might stretch our data thin.





Direction & Next Steps

In Progress

Sequential Decision-Making Experiment

- Run over a sequence of time in historical data, iteratively selecting cases and training on the results.
- Comparison of baseline vs. imputation vs. meta-learning strategies and model architectures.
- Early results show that simply increasing the cadence of updates from several years (as in NRP) to single years (and further to monthly) already yields a performance lift regardless of the use of intermediate outcomes.

Identifying Sub-Populations

- Leveraging intermediate outcomes may be more (or less) useful in certain case contexts. Can we characterize these? Does it make sense to focus on (or exclude) certain areas?

Future

Scenario Exploration

- Under what conditions do the conclusions of our experiments hold? Conditions could include volume of data, complexity of inference space, levels of drift / response to drift events.

Integration with Detection-controlled Estimation (DCE)

- Intermediate outcomes may mean different things depending on where they come from. We might try to integrate detection-controlled estimation to condition this out from our insights.

Resource Management

- The meta-learning approach potentially yields information that could be used for casework decision-making. What are the opportunities and pitfalls here?