Does Explainable Machine Learning *Really Matter?*

AKA: How I Learned to Stop Worrying and Love the Bayes with rstanarm

AKA: 2 Fast 2 Machine Learning: Tokyo Model Drift

#justice4Han

Selina Carter
Carnegie Mellon University
@selina_carter_

[Part1!: https://youtu.be/fWfSGI-pf0A](https://youtu.be/fWfSGI-pf0A)

AKA: Dammit Jared Is Always Right

Jonathan Hersh
Argyros School of Business
Chapman University
@DogmaticPrior

Why Do We Care About Explainable AI?

1. Consequential decisions may need human validation
2. Explaining the model may help us build better models

Opinion

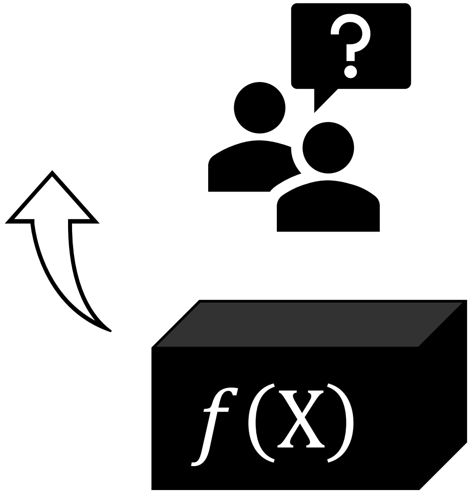
OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

230



BUSINESS | HEALTH CARE | HEALTH

Researchers Find Racial Bias in Hospital Algorithm

Healthier white patients were ranked the same as sicker black patients, according to study published in the journal Science

An algorithm widely used in hospitals to steer care prioritizes patients according to health-care spending, resulting in a bias against black patients, a study found.

PHOTO: GETTY IMAGES

By Melanie Evans and Anna Wilde Mathews

Updated Oct. 25, 2019 8:39 am ET

Our Machine Learning Model: Predicting Delays in Loans

- We partnered with a major development bank to build ML models to predict delays for sovereign guaranteed investment loans.
  - Think: large infrastructure loans
  - Avg size $67m USD.
  - Delays costly: 22% loans delayed, 24% of supervision cost from delays

RG-L1124 :
Modernization of the Salto Grande Binational Hydropower Complex

Project Status: Implementation

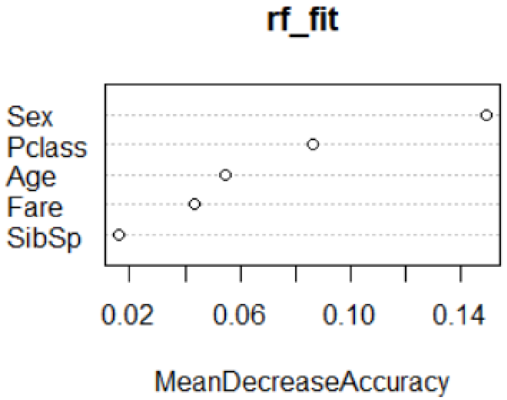
The overall objective is to help ensure the availability of the Salto Grande Hydropower Complex(SGHC), enhancing the reliability and efficiency of the interconnection between Argentina and Uruguay. The specific objective is to assist in extending the useful life of the SGHC by modernizing its infrastructure and equipment

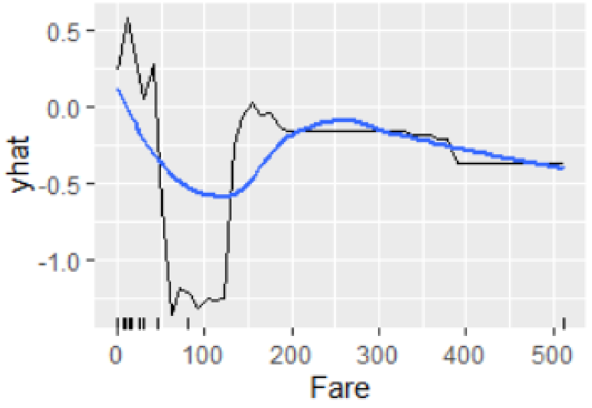
PROJECT INFORMATION

TOTAL COST	USD 80,000,000
COUNTRY COUNTERPART FINANCING	USD 0
AMOUNT	USD 80,000,000

What Kind of Explainable AI Exists?

- Global feature importance and marginal effects (partial dependence plots)

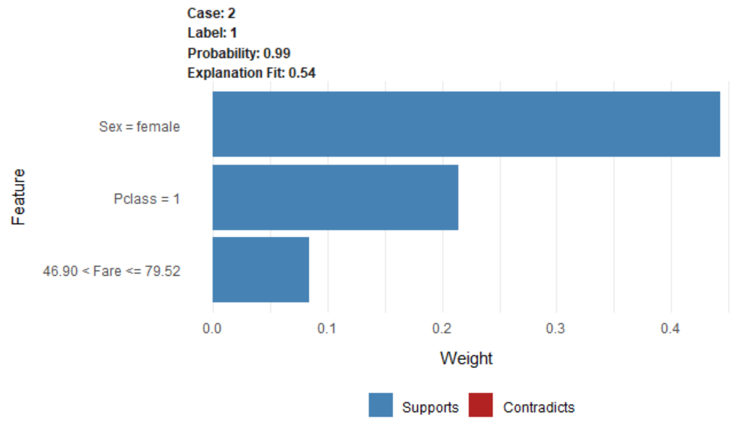


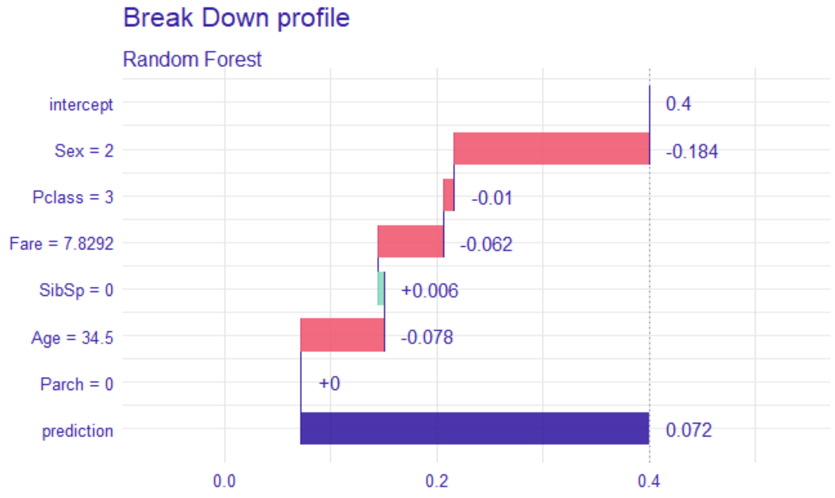


Maybe our models are too complicated? Angelino, Larus-Stone, Alabi, Seltzer, and Rudin. **Learning Certifiably Optimal Rule Lists for Categorical Data**. JMLR, 2018.

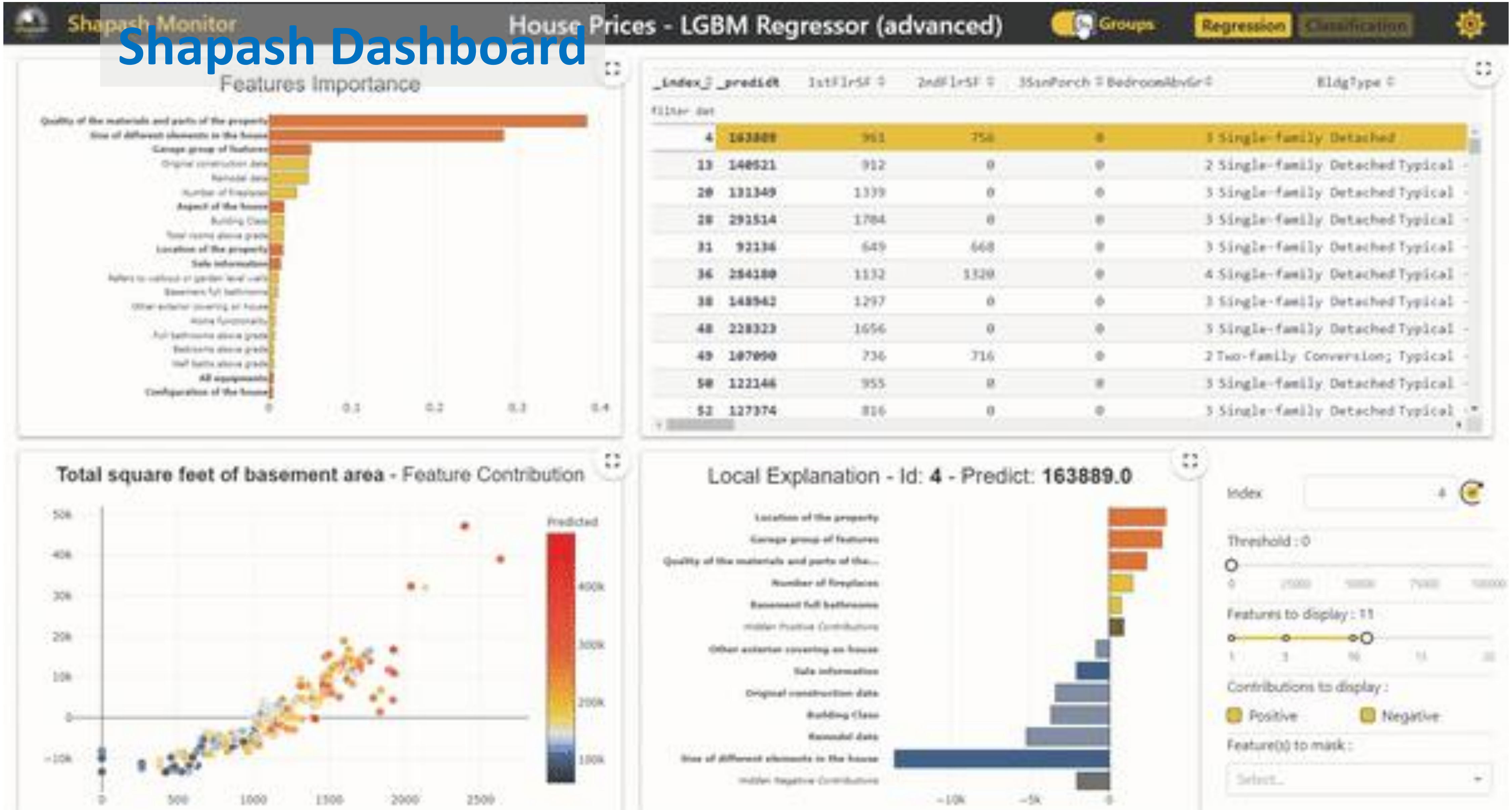
https://github.com/corels/rcppcorels

- Explaining individual predictions (LIME) and Shapley values



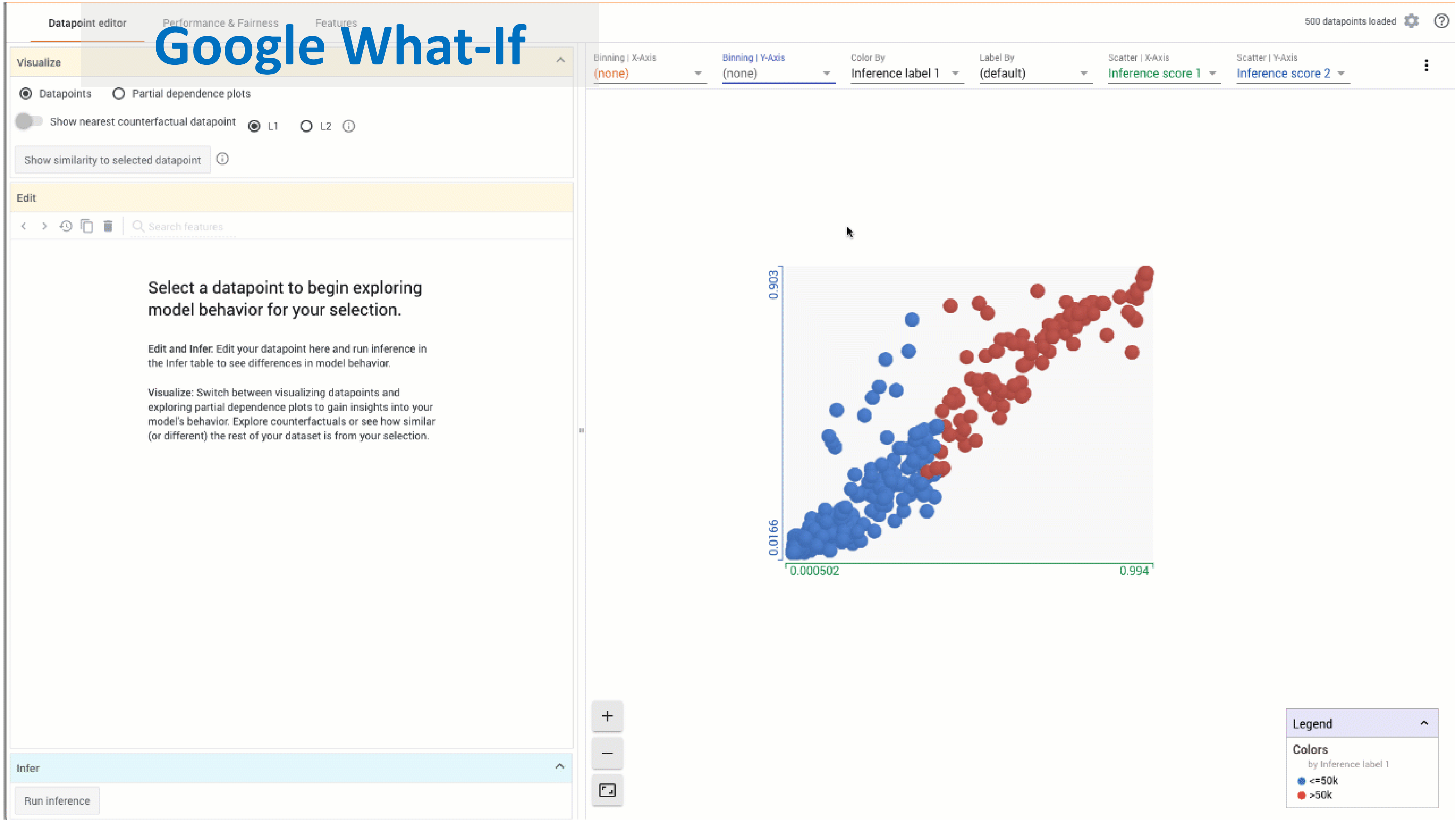


Shapash Dashboard



https://github.com/MAIF/shapash

Google What-If

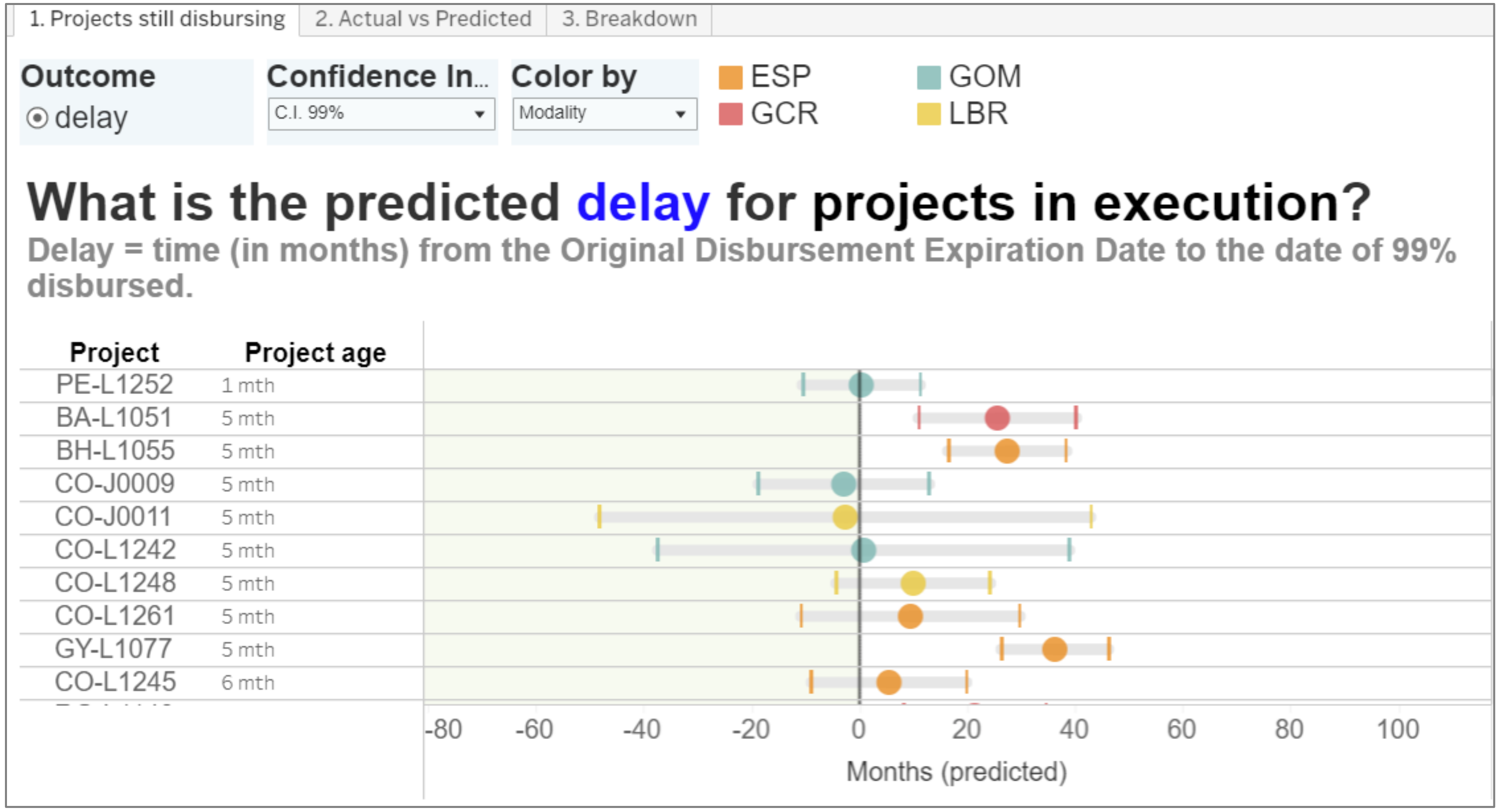


https://pair-code.github.io/what-if-tool/

Dashboard and Survey

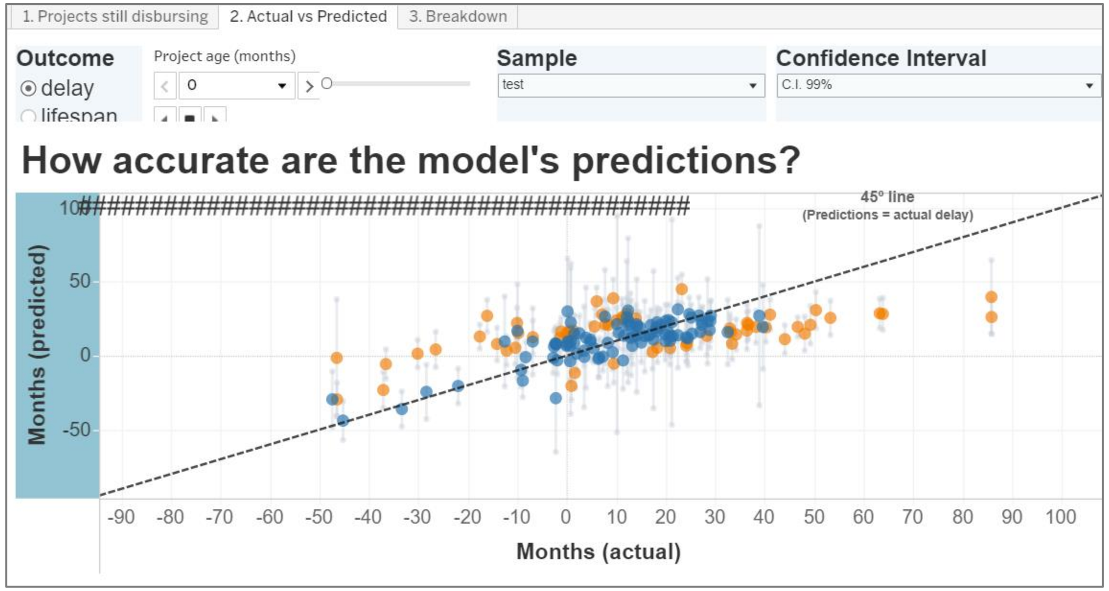
- **Sample:** 685 team members (managers, analysts)
- **Timeline:** May-June, 2021
- **Responded**: 617 (90%)

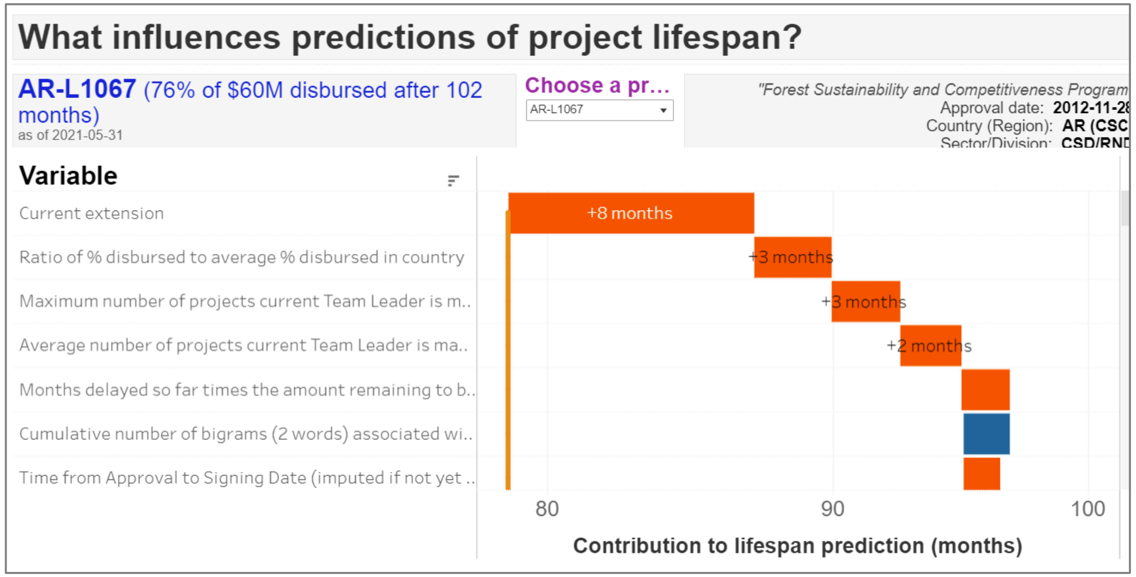
- **Control group**: $(N_{control} = 263)$
  - Just predictions and confidence intervals



Explainability/Model Performance Treatment Group

- i) Predictions and CIs
- ii) **information on model performance**
- iii) **Model explanation** describing why each project was given a particular delay (LIME and Shapely plots)
- $N_{Treatment} = 227$





Results Analysis

- **Four outcome variables:**
  - How useful is the ML tool? (Ordered rank 1-5)
  - How well did you understand the ML tool? (Ordered rank 1-5)
  - Did user change their delay estimate before/after viewing tool (=1 if updated)
  - Absolute value of change in delay estimate update

- **Five Moderators** (Pre-registered AsPredicted.org #69245)**:**
  - Explainable model/performance dashboard
  - Work Location
  - Role (Team Leader, Analyst, Fiduciary/Procurement, Chief of Operations)
  - Loan Amount
  - Machine Learning Familiarity (1 least, 5 most)

- **Ordered Logistic, Poisson, or Logistic Depending on Outcome**

Jonathan Hersh @DogmaticPrior · Jul 21

What are some good #rstats resources/packages for ordinal multinomial regression? I've never really used these before but need to crack em open for a project.

4 1

Jared Lander @jaredlander · Jul 21

Replying to @DogmaticPrior

rstanarm*

1 1

Grant McDermott @grant_mcdermott · Jul 21

Replying to @DogmaticPrior

cran.r-project.org/web/packages/m...

(But I agree with @abiylfoyp that this is often a great use case for Bayes.)

mlogit: Multinomial Logit Models
Maximum likelihood estimation of random utility discrete choice models. The software is ...
cran.r-project.org

1 1

Yondren McDarry @abiylfoyp · Jul 21

The data I spent a lot of time with is likert with many observations per rater. In that setting, some raters are Uber-style always-5-stars; others are 1 OR 5, and others grade more carefully--the common-cut-point assumption is a bad one.

1 2

Show replies

Jesse Bruhn @jmb112485 · Jul 21

Replying to @DogmaticPrior

lm(y~x)

1

Why you fear going Bayes

Plain Laziness!

What Priors to Use?

Critical Referees!

(c) Using the flat prior of $f(p_1, p_0) = 1$, derive the joint posterior density of (p_1, p_0). Then, via simulation, find the posterior mean and 90% central interval for τ.

Solution. First calculate likelihood for $p1, p0$:

$$\begin{aligned}
L(p_1, p_0 \mid Y) &= \prod^{N} f(Y \mid p_1, p_0) = \prod^{N_1} f(Y_1 \mid p_1) \prod^{N_0} f(Y_0 \mid p_0) \\
&= \prod^{N_1} f(Y_1 \mid p_1) \prod^{N_0} f(Y_0 \mid p_0) \\
&= \left[\prod^{N_1} p_1^{y_{i1}} (1 - p_1)^{1 - y_{i1}}\right] \left[\prod^{N_0} p_0^{y_{i0}} (1 - p_0)^{1 - y_{i0}}\right] \\
&= p_1^{\sum y_{i1}} (1 - p_1)^{N_1 - \sum y_{i1}} p_0^{\sum y_{i0}} (1 - p_0)^{N_0 - \sum y_{i0}}
\end{aligned}$$

And then the joint posterior density is:

$$\begin{aligned}
P(p_1, p_0 \mid Y) &= L(p_1, p_0 \mid Y) f(p_1, p_0) \\
&= \left[p_1^{\sum y_{i1}} (1 - p_1)^{N_1 - \sum y_{i1}} p_0^{\sum y_{i0}} (1 - p_0)^{N_0 - \sum y_{i0}}\right] \quad (1)
\end{aligned}$$

JAGs? Stan? WinBugs?

- R Code- -

```
Init <- list(list(a = 0, b = 0,
     .RNG.seed = 1234,
     .RNG.name = "base::Mersenne-Twister"),
     list(a = 10, b = -10,
     .RNG.seed = 5678, .RNG.name = "base::Super-Duper") )
j.out <- jags(data = Data, inits = Init,
     parameters = c("a", "b"),
     model = "logit.jag", n.thin = 5,
      n.iter = 10000, n.chains = 2)
c.out <- as.mcmc(j.out)
```

Jonathan Hersh @DogmaticPrior · Jul 21

What are some good #rstats resources/packages for ordinal multinomial regression? I've never really used these before but need to crack em open for a project.

4 1

Show more replies

Jared Lander @jaredlander · Jul 21

Replying to @DogmaticPrior

rstanarm*

1 1

Jonathan Hersh @DogmaticPrior · Jul 21

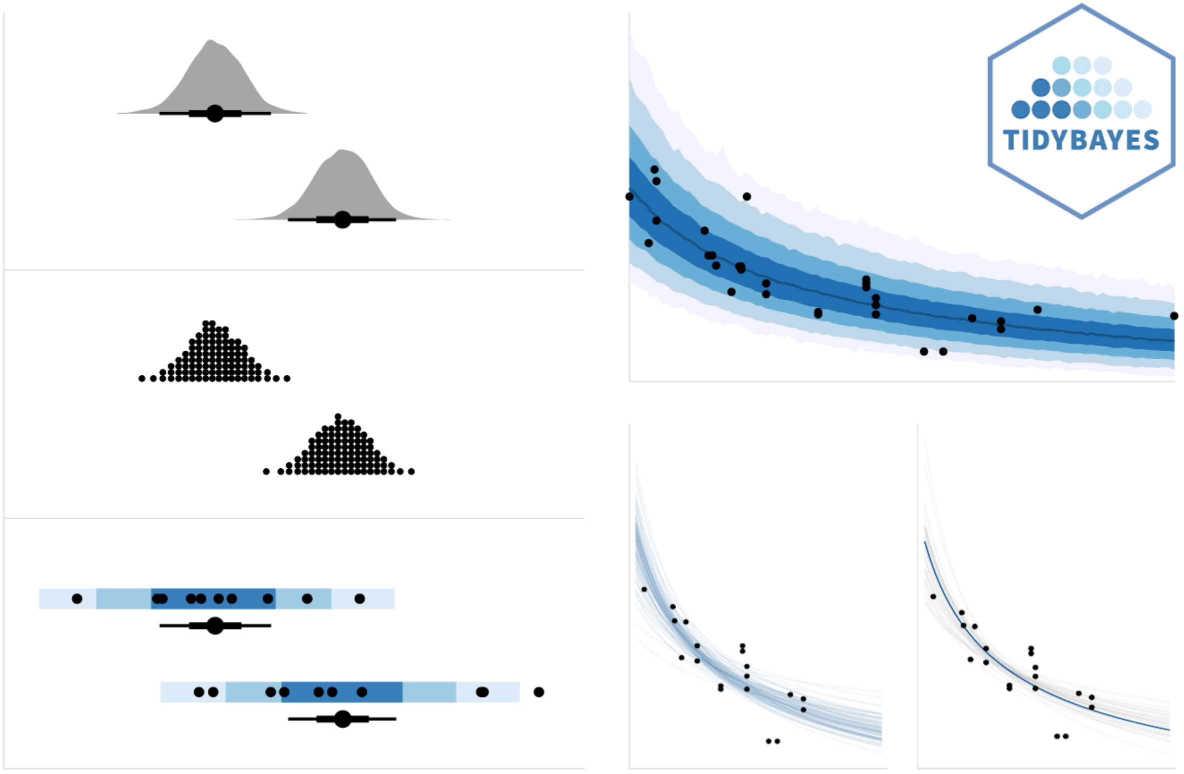
Honestly, this is the right answer.

Bayes just feels way more honest.

mc-stan.org/rstanarm/artic...

1

Why have I slept on you, rstanarm, TidyBayes & friends?

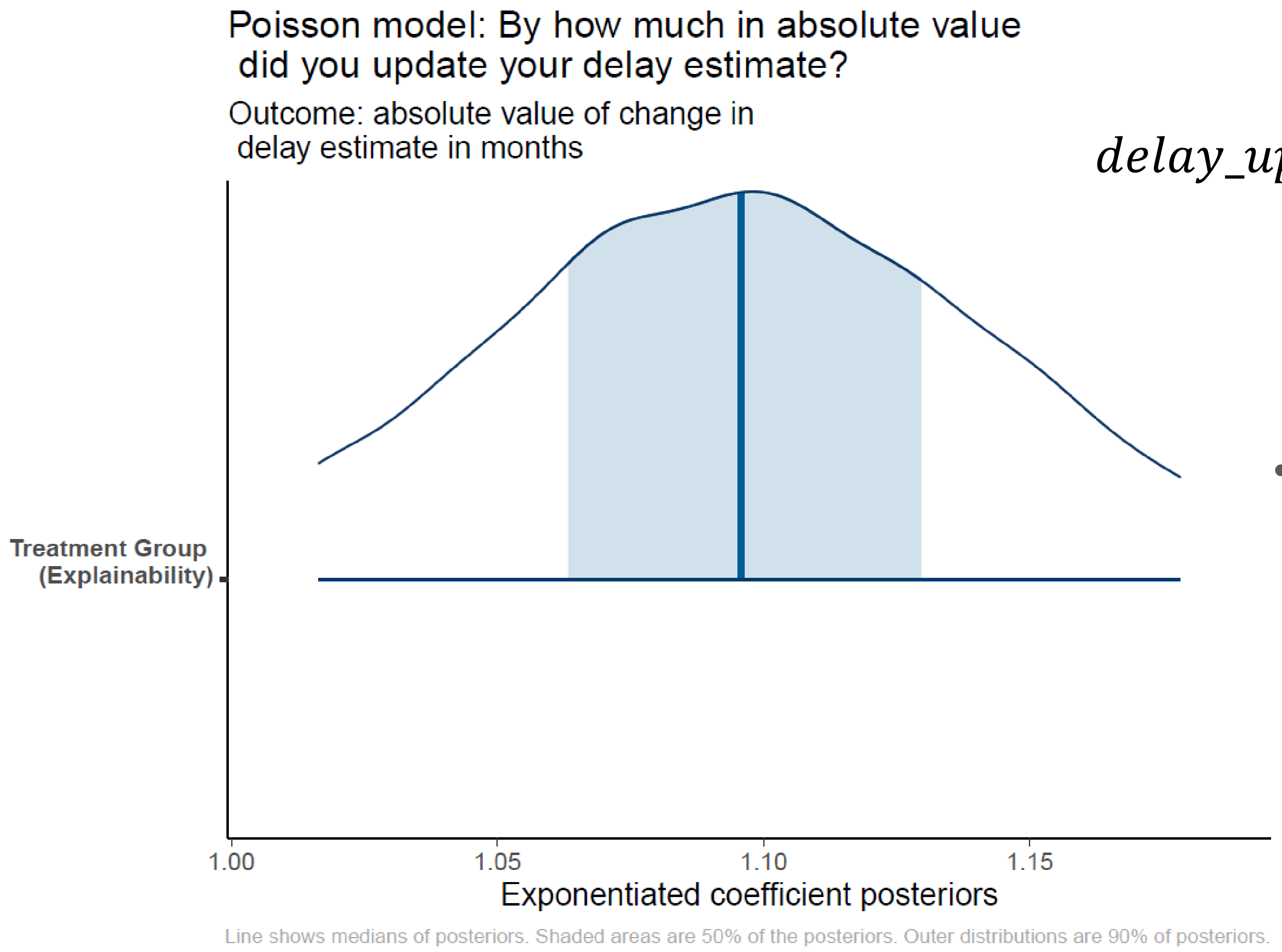


rstanarm

rstanarm is an R package that emulates other R model-fitting functions but uses Stan (via the rstan package) for the back-end estimation. The primary target audience is people who would be open to Bayesian inference if using Bayesian software were easier but would use frequentist software otherwise.

Fitting models with **rstanarm** is also useful for experienced Bayesian software users who want to take advantage of the pre-compiled Stan programs that are written by Stan developers and carefully implemented to prioritize numerical stability and the avoidance of sampling problems.

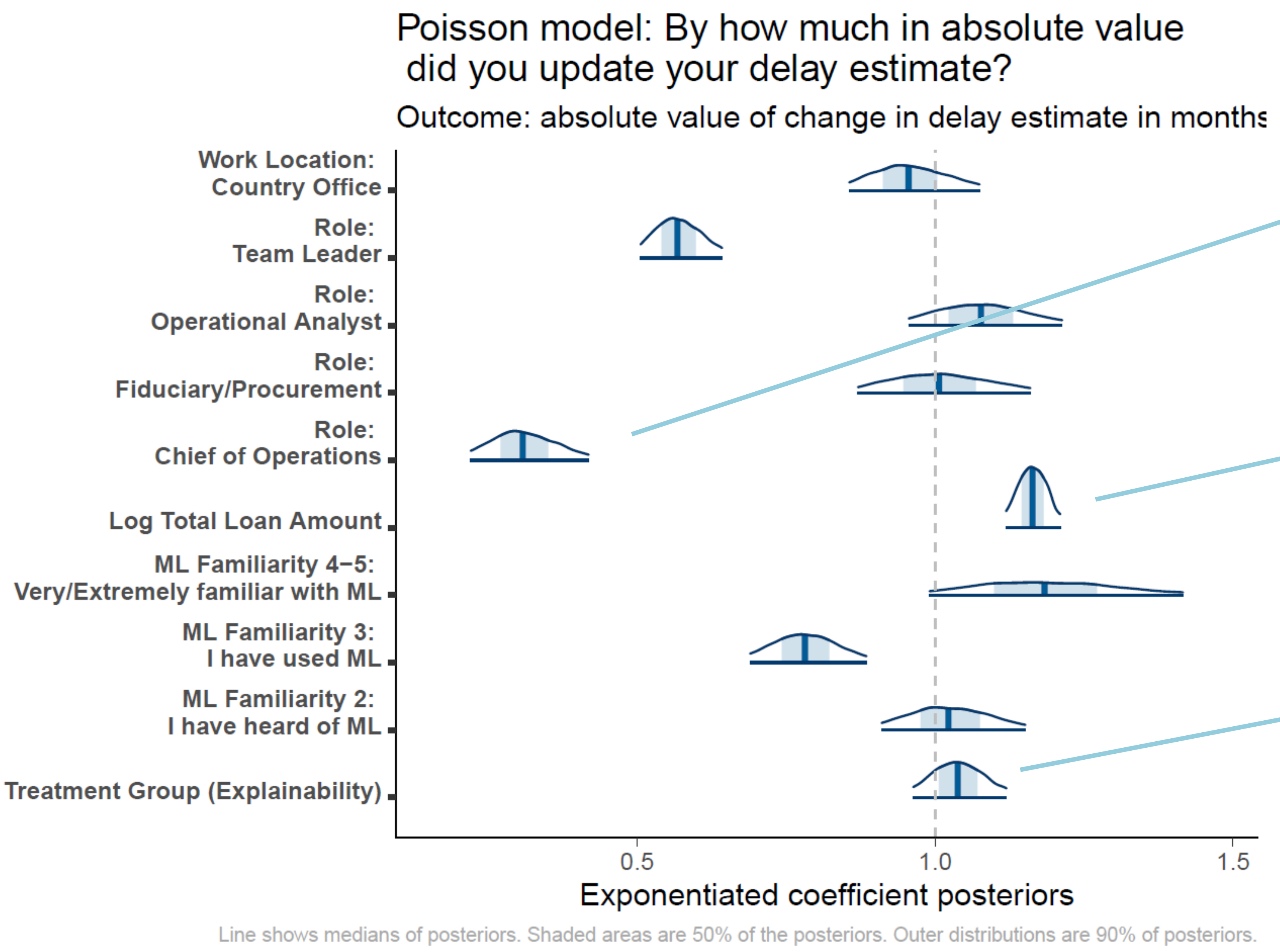
http://mjskay.github.io/tidybayes/



$$delay\_update_i = \exp(x_i' * \beta_{Treatment})$$

- Median: $\exp(\beta_{Treatment}) = 1.1$

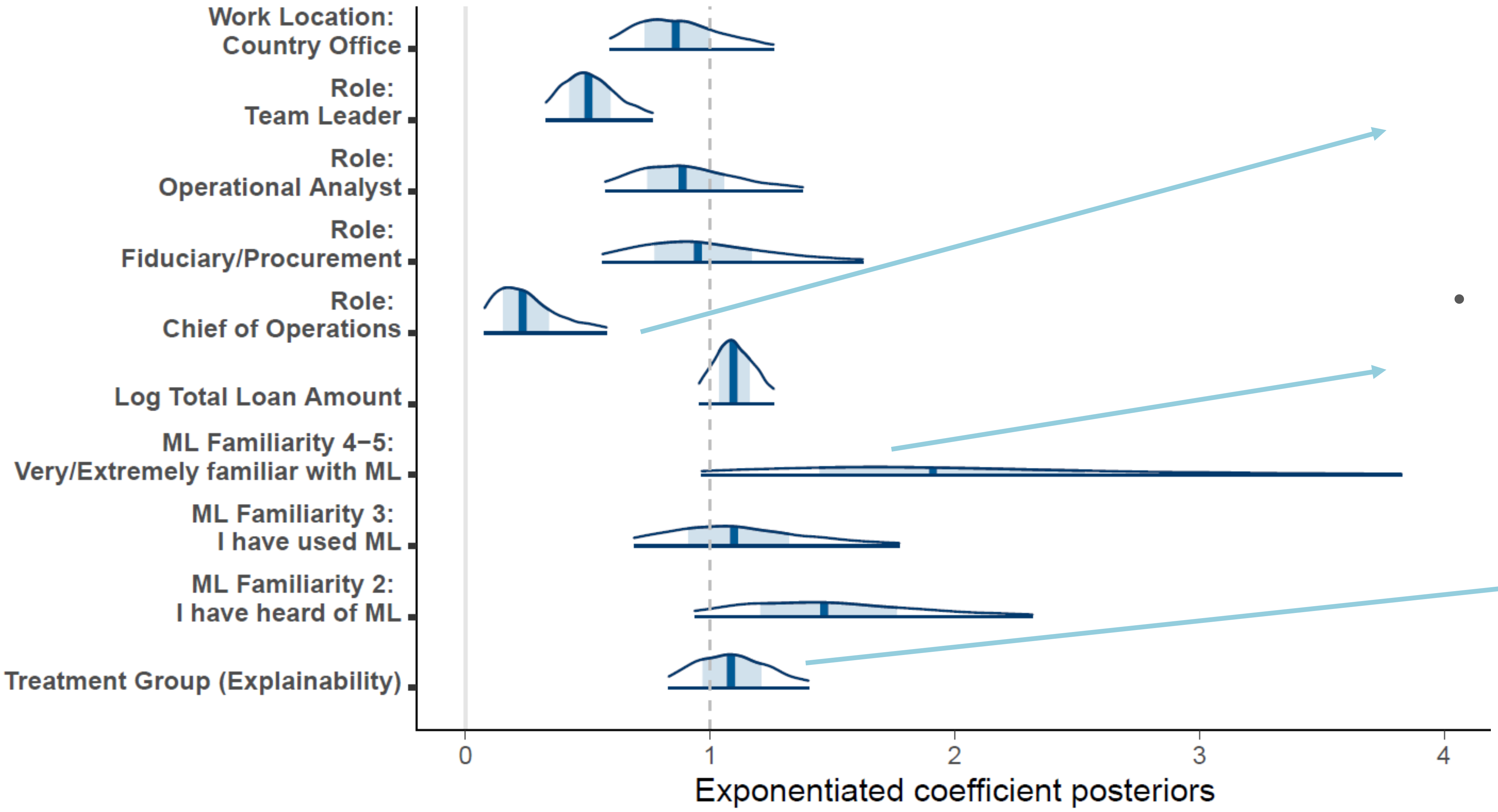
  Average delay update = 2.2 months -> 4.5% impact of explainability treatment



- Most senior member of team updates beliefs by lowest amount
- Higher loan amount -> more likely to update beliefs
- After controlling for other individual characteristics, treatment effect is mixed

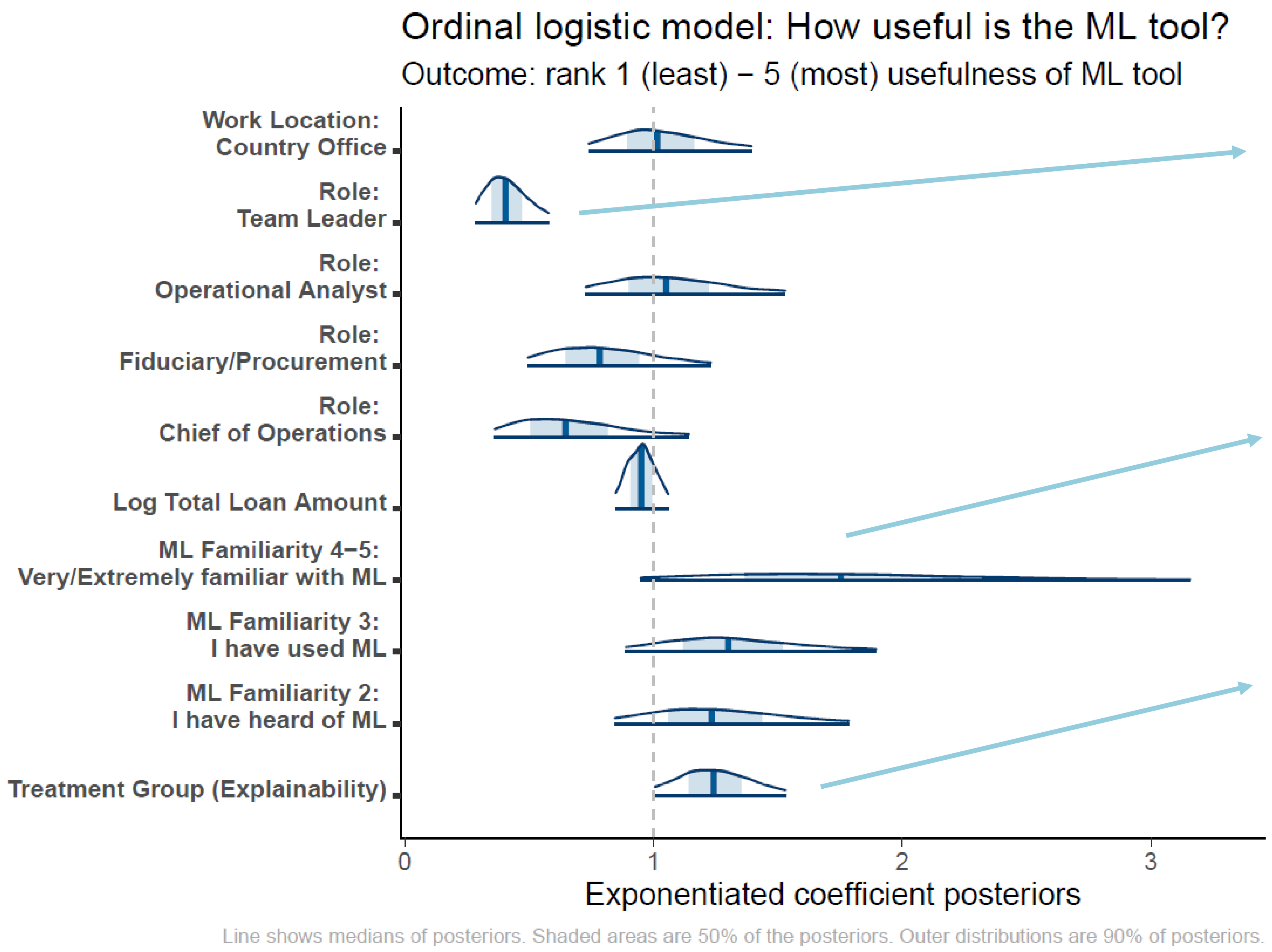
Logistic model: After viewing the ML tool did you update your delay estimate?

Outcome: =1 if changed delay estimate for project after viewing ML tool



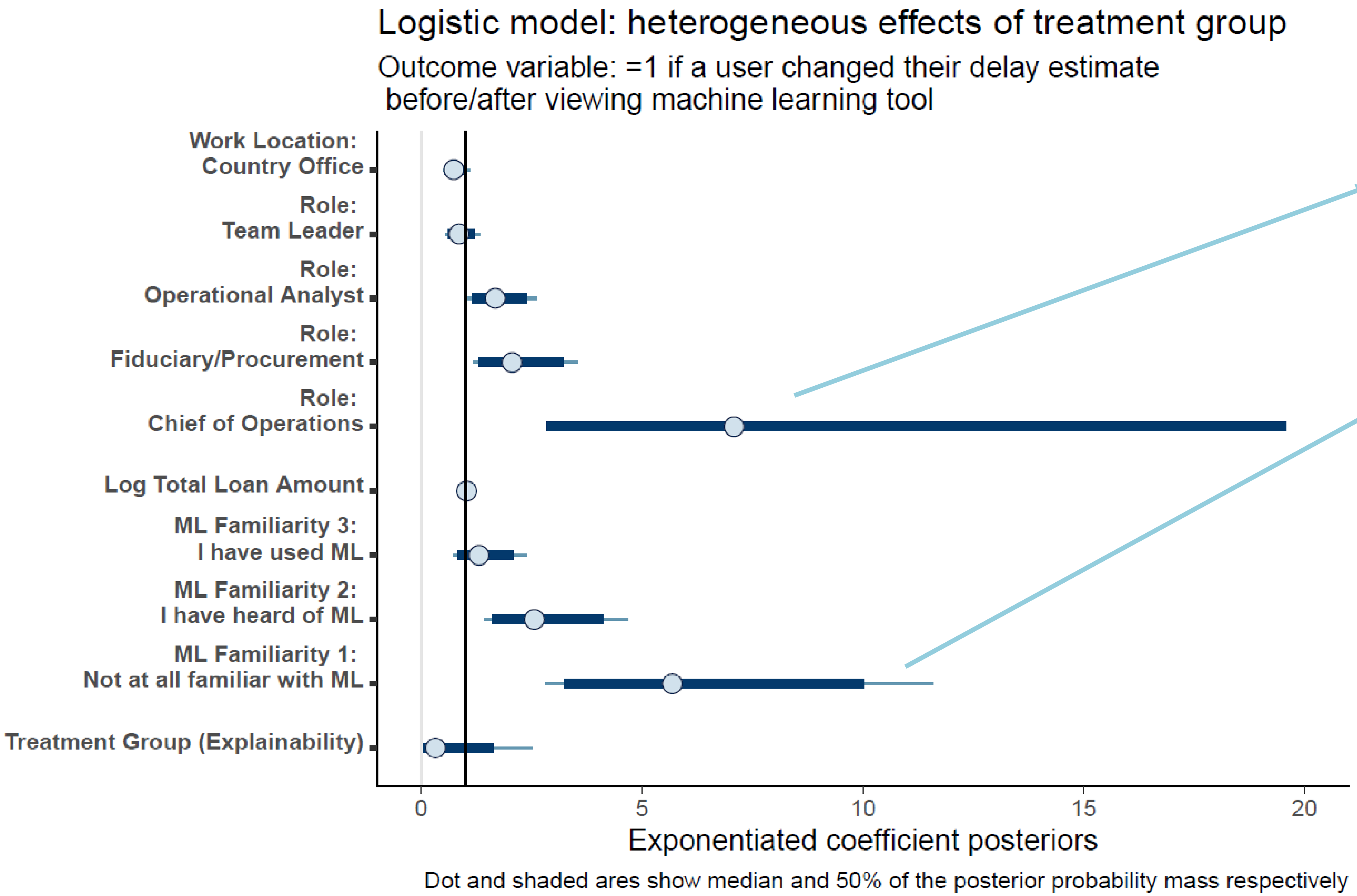
Line shows medians of posteriors. Shaded areas are 50% of the posteriors. Outer distributions are 90% of posteriors.

- Median: $\exp(\beta_{Chief}) = 0.23$.
  - Most senior member of team 77% less likely to update delay estimate!
- Most familiar with ML 2x as likely to update delay estimate
- Again, appear to be mixed impact of explainability treatment on belief update



- Direct managers least likely to find the tool useful
- Users with higher machine learning knowledge more likely to find ML tool useful
- Median: $\exp(\beta_{Treatment}) = 1.24$ ~= 24% increase in likelihood of increasing self reported usefulness if receive explainability treatment

Which groups are most likely to respond to explainable AI?



- Least familiar with ML and most senior members respond most to explainability treatment

- Heterogeneous treatment effects on these groups are very large. 5x-7x more likely to update belief if in explainability treatment

Conclusions

- We find that explainable AI models increases belief updating by 4.5%

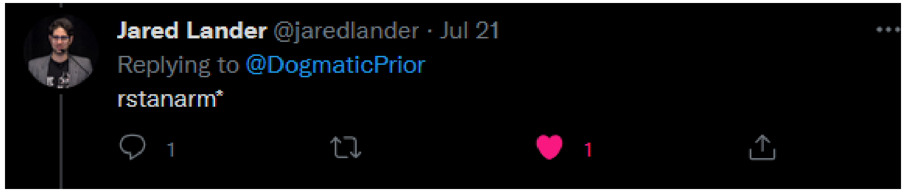
- Explainable models increase perceived usefulness but decrease understanding

- Largest loans more likely to update beliefs given ML predictions

- Senior members of team and those least familiar with ML least like to trust AI

    - But: explainable models increase belief updating by these reticent groups by 5-7x.

WHO IS THE CAPED CRUSADER?



https://tinyurl.com/superjared

Thank you
Fiverr artist Nuril Anwar

https://tinyurl.com/superjared

Comments, Questions or Suggestions

People

- Selina Carter, Carnegie Mellon University – selinahowcarter@gmail.com
- Jonathan Hersh, Argyros School of Business and Economics, Chapman University – hersh@chapman.edu