Old Methods with New Tricks

Our Solution for the PrecisionFDA
Brain Cancer Predictive Modeling & Biomarker Discovery Challenge

Nan Xiao, Soner Koc, Kaushik Ghose

Seven Bridges

Agenda

- Results
- Solution
  - Feature selection
  - Stability selection
  - Predictive modeling
- Lessons

Our Results

All 4 clinical features are used without selection in all sub-challenges

Sub-challenge 1

- 16 markers: CAPN14, COX8C, FILIP1, IQCF5, LACTB2.AS1, LINC01537, SLC13A1, ZNF407, C1QTNF7, LINC00635, LINC00691, PPP4R3C, NXNL2, FARSB, RGS13, CPXCR1
- AUC on training (fitting): 100%

Sub-challenge 2

- 6 markers: 3q13.12, 3p11.1, 9p12, 21q11.1, 17q23.1, 20p11.1
- AUC on training (fitting): 91.19%

Sub-challenge 3

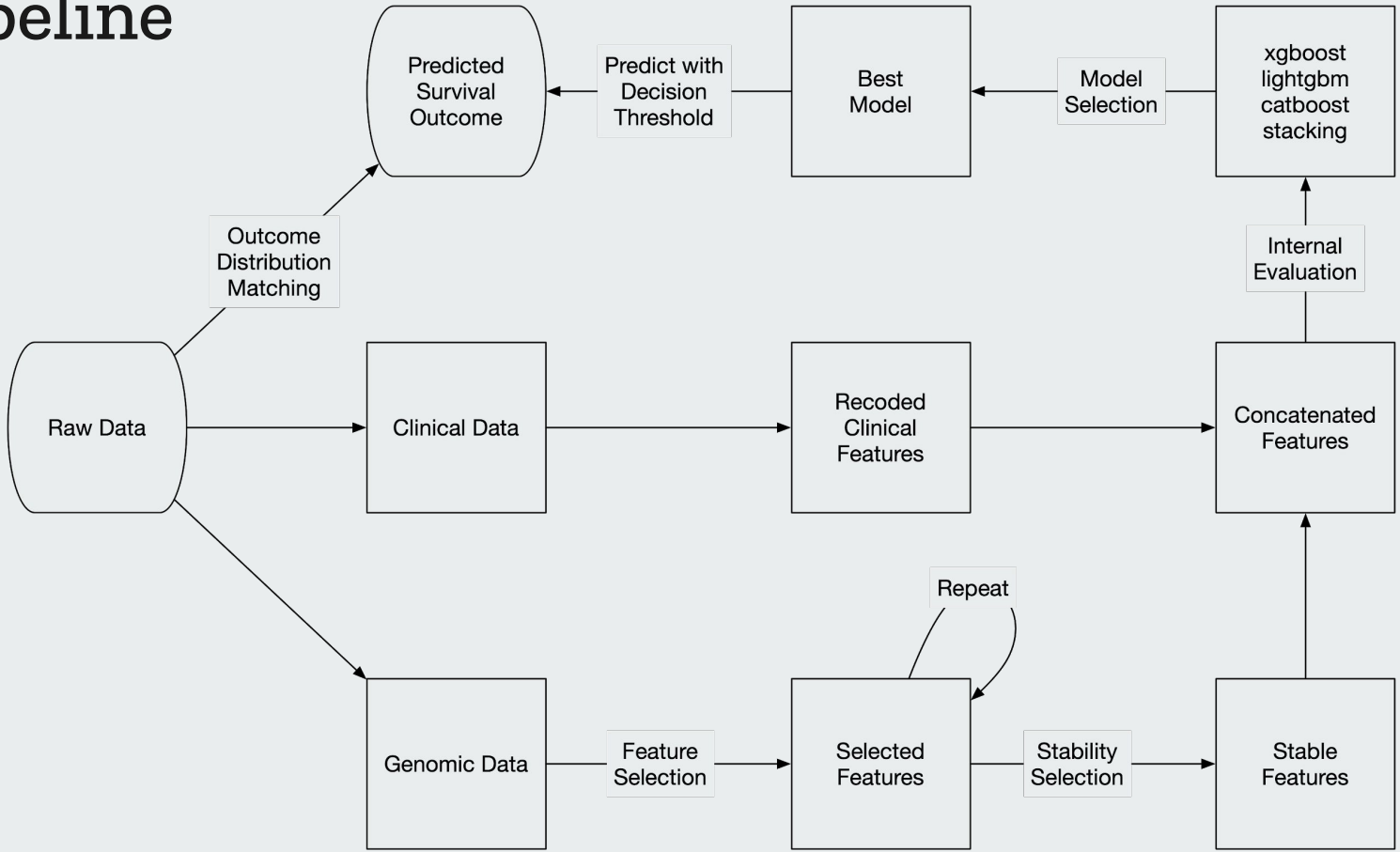
- 13 markers: C17orf77, HELQ, LOC101927870, MAGEB18, TRAV12.3, ADH4, AVPR1A, PDE6C, LINC01854, BRD7P3, LINC01085, CT83, LINC01815
- AUC on training (fitting): 100%

Our Solution

Principles

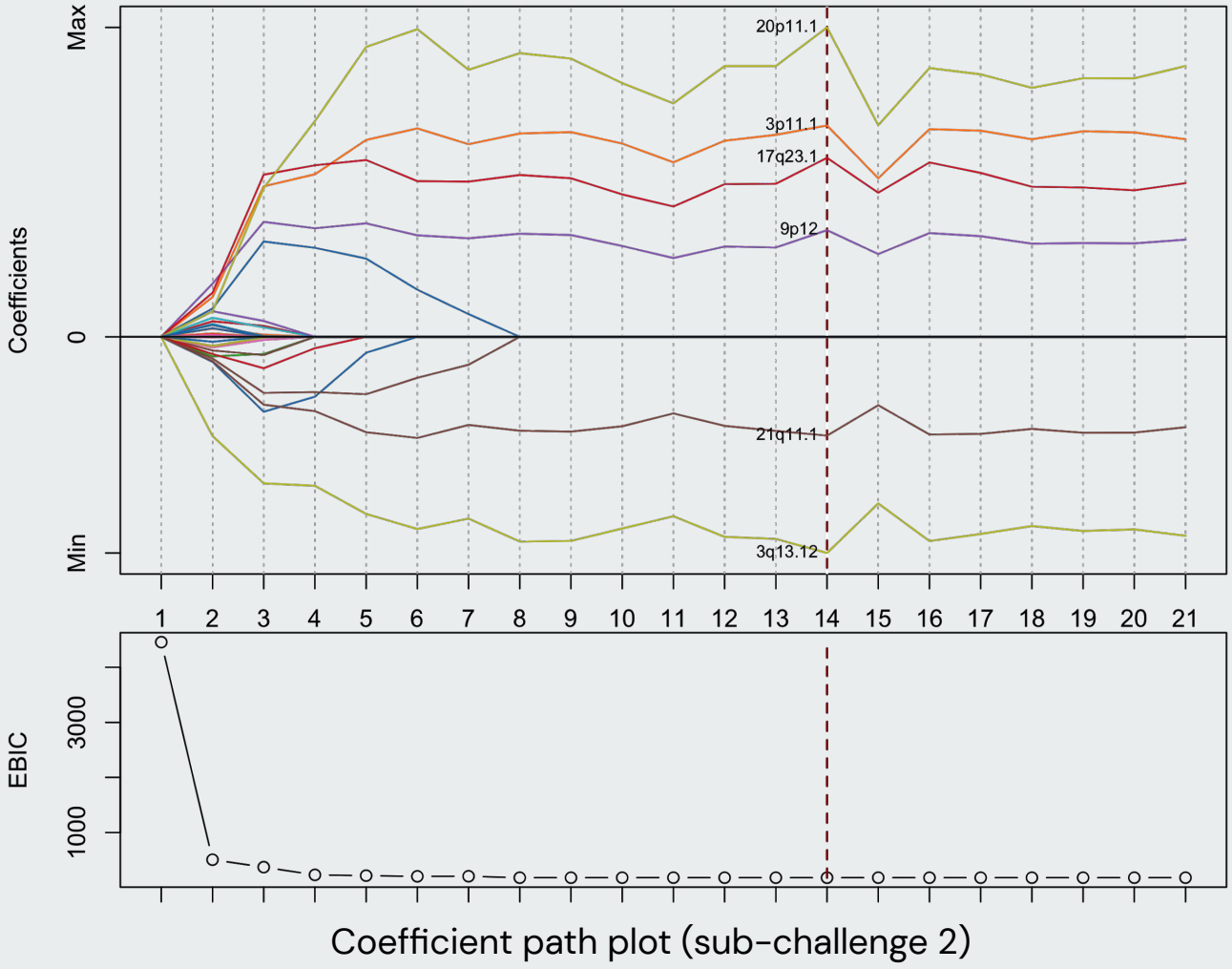
1. One unified, automated pipeline but allows human insights
2. Selected genomic features + all clinical features
3. Highly sparse solutions for feature selection (<20)
4. Ensemble of ensemble models for predictive modeling
5. Tackle class imbalance in prediction (inference) phase

Pipeline


Predicted Survival Outcome
Predict with Decision Threshold
Best Model
Model Selection
xgboost
lightgbm
catboost
stacking
Outcome Distribution Matching
Internal Evaluation
Raw Data
Clinical Data
Recoded Clinical Features
Concatenated Features
Repeat
Genomic Data
Feature Selection
Selected Features
Stability Selection
Stable Features

1. Feature Selection

- Model-based feature selection: sparse logistic regressions
- Multi-step adaptive SCAD-Net implemented by `msaenet`
- Adaptive elastic-net for sparse regressions (Zou & Zhang, 2009)
- Multi-step adaptive estimation can reduce the # of false positive selections while maintaining predictive (Xiao & Xu, 2015)
- Nonconvex penalties such as SCAD (Fan & Li 2001) and MC+ (Zhang, 2010) can further reduce bias in each estimation step



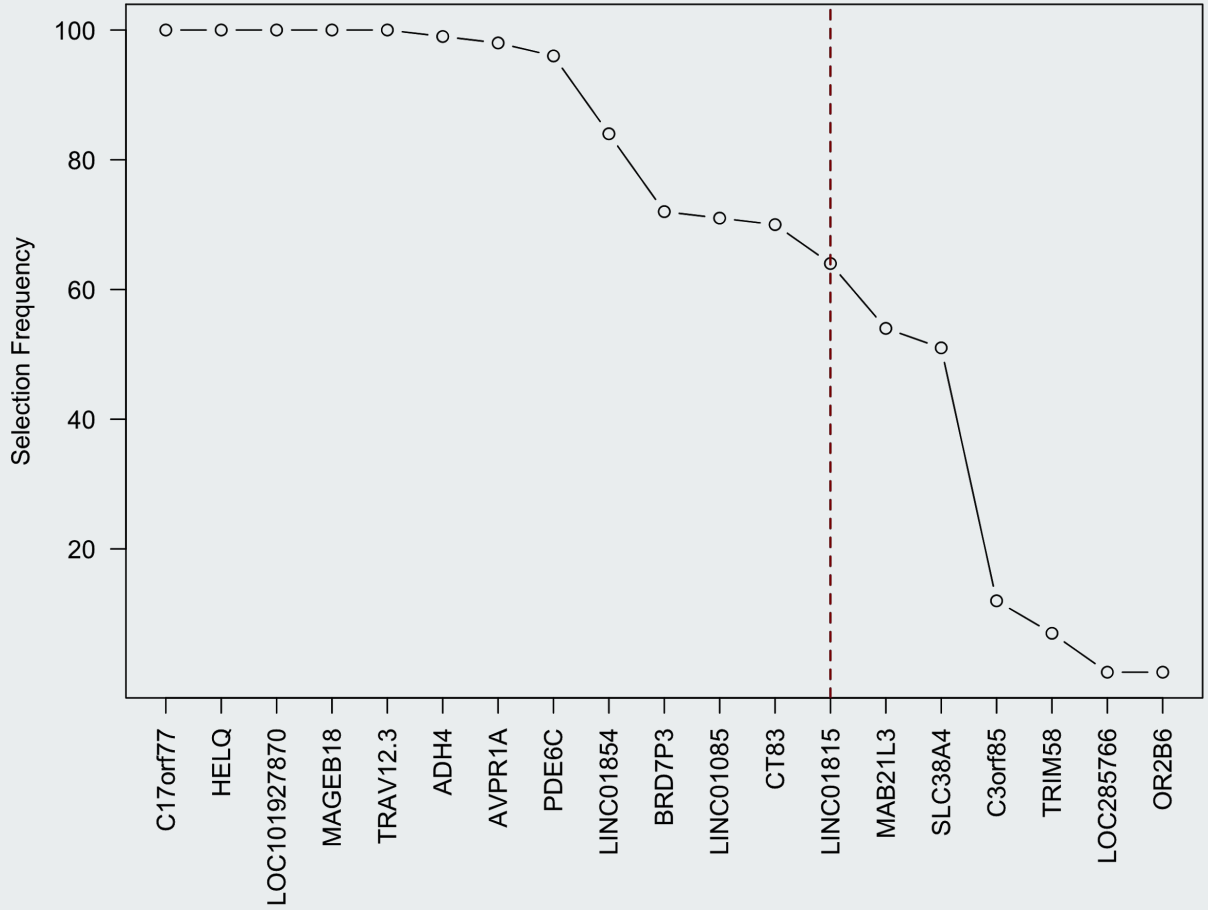
Coefficient path plot (sub-challenge 2)

Feature Selection Summary

- Trick: multi-step adaptive estimation + nonconvex penalties
- Experiments with Lasso and LOLearn penalties
- No selection on clinical features: let the tree models work

2. Stability Selection

- For selective inference (Meinshausen & Bühlmann, 2010)
- Resample the dataset and aggregate feature selection results from many models
- We borrowed the idea and used a relaxed version to eliminate the “unstable” features



Relaxed stability selection (randomization + aggregation) for sub-challenge 3

Stability Selection Summary

- Trick: randomization + aggregation to distill stable features
- Introduce human decisions at the right moment

3. Predictive Modeling

- GBDT with xgboost, lightgbm, and catboost
- Stacking ensemble (Wolpert, 1992) of the three models
- Only tuned the three most important parameters
  - max tree depth
  - learning rate
  - Iterations

Predictive Modeling Summary

- Trick: stacking ensembles + decision threshold tuning
- Stacking ensembles work even with similar base learners
- Pay extra attention to your model stacking code
- Tune decision threshold to match the training set class distribution. Useful for combating class imbalance.

Lessons Learned

Reflections on Performance Evaluation

- External comparison

  Our models are comparably smaller --- lower cost in experimental validation and productization. Metrics like Extended BIC are helpful.

- Internal comparison

  We did better in the more difficult tasks (2 and 3). Should look into task 1 given the chance – maybe too sparse.

- Submission process

  No public leaderboard thus no feedback. Use multiple solutions to diversify submissions and reduce risk.

Software availability

msaenet

https://github.com/nanxstats/msaenet

stackgbm

https://github.com/nanxstats/stackgbm

Pipeline

https://github.com/nanxstats/bcpm-msaenet