Predict survival for cancer patients using real world data: simple model or advanced machine learning?

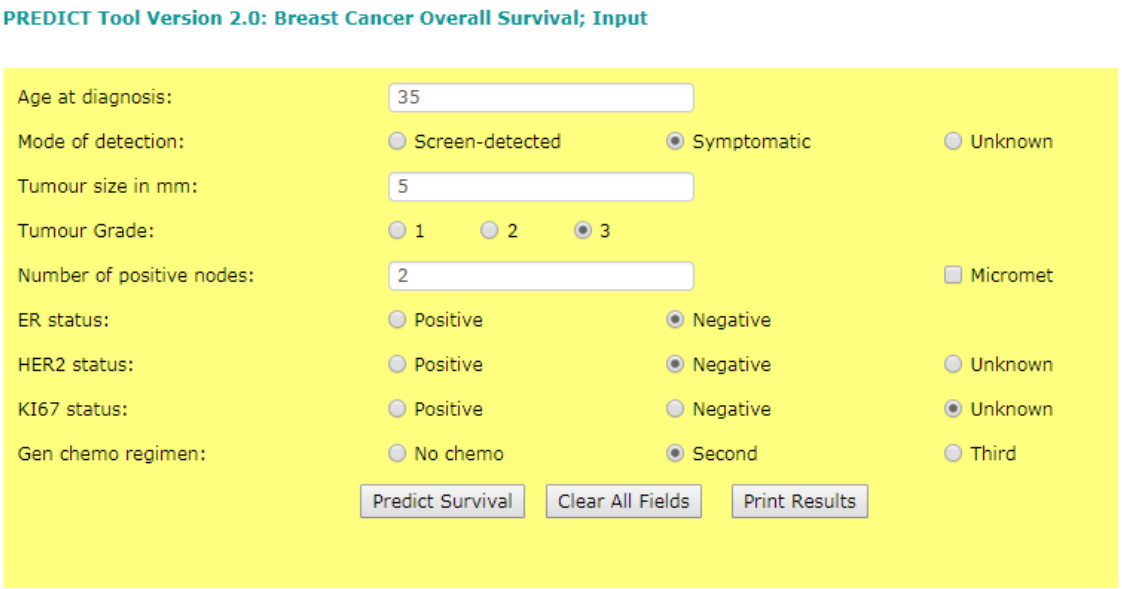
Guiyuan Lei, Saeed Rafii and Kenji Hashimoto
Roche Products Ltd
PSI 2018 conference, 4th June 2018, Amsterdam, Netherlands

Outline

- Commonly used tools/methods for survival prediction
- One real example: Roche RAAD Challenge
  - The advantage of the data and challenge of the problem
  - Approaches from different teams
  - Comparing simple model vs advanced machine learning
- Summary/Learnings

Commonly used tools/methods for survival prediction

- Two well-known survival prediction tool for eBC patients (Cox-regression model, meta-analyses)
  - Adjuvant! Online
  - Predict



- Commonly used Machine Learning Methods for survival prediction
  - Artificial Neural Networks (ANNs)
  - Support Vector Machines (SVMs)
  - Bayesian Networks (BNs)
  - Decision Trees (DTs)
  - …

One real example: Roche RAAD Challenge

RAAD Challenge:
The advantage of the data and challenge of the problem

The Objective of the RAAD Challenge

The Roche Advanced Analytics Data (RAAD) Challenge was an internal competition within the Roche Group to predict the probability a patient will be alive at 1 year after treatment initiation, using Flatiron electronic health record data.

Flatiron Data

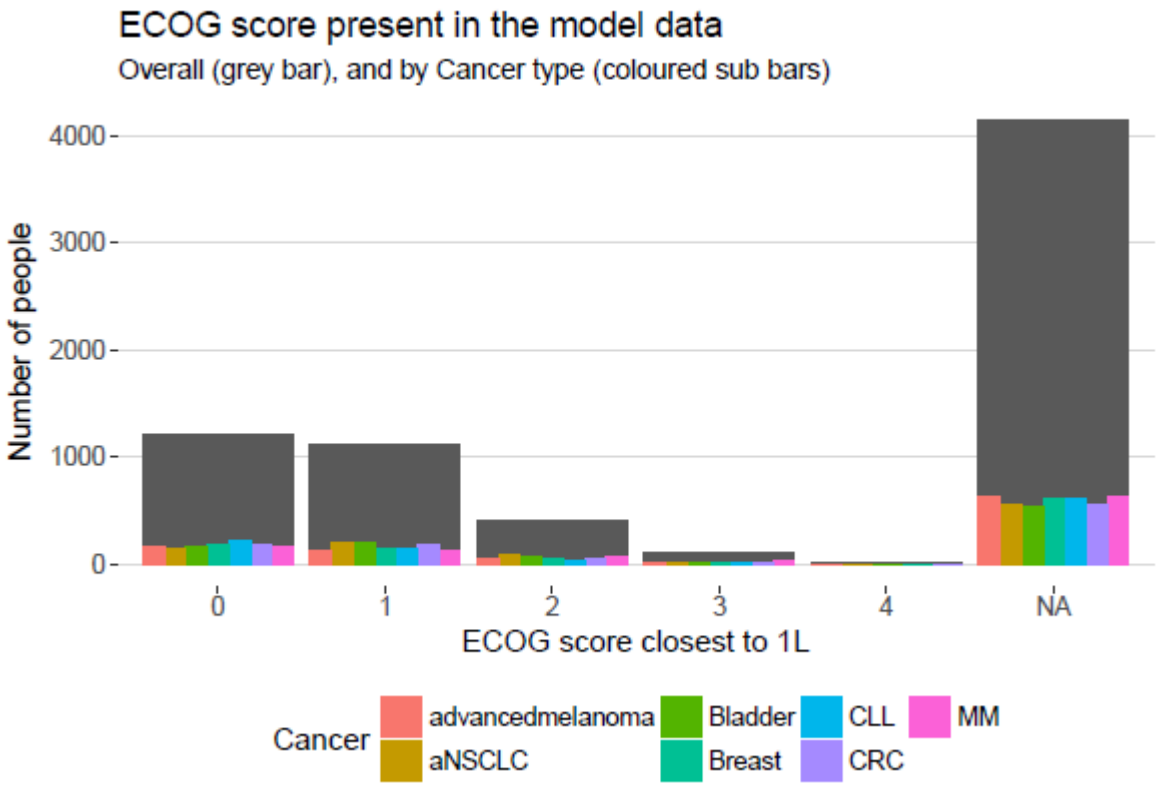
- The Flatiron Health database is a longitudinal, demographically and geographically diverse database derived from electronic health record (EHR) data
- The database includes data from over 265 cancer clinics (~800 sites of care) representing more than 2 million active US cancer patients available for analysis
- This challenge utilized a random sample of the Flatiron Health EHR-derived database and included 10500 patients (7000 for training and 3500 for testing) diagnosed with 7 different cancer types. The treatment information was not included in challenge.
- Patient-level data include structured and unstructured data, curated via technology-enabled abstraction
- Data provided are de-identified and provisions are in place to prevent re-identification in order to protect patients' confidentiality

The advantage of the data and challenge of the problem

- Advantages of the data
  - There were 20 datasets including **generic tables** and **indication specific tables** (demo, biomarker, metastatic information etc for 7 different cancer types)
  - Wealthy data, for example, there were ~120 lab variables/parameters, some of them have multiple time points (longitudinal data)

- Challenge of the problem:
  - Data is imbalanced/skewed in regards to number of patients dead or alive. For CLL, only 10% patients have died within one year.
  - The percentage of missing data is higher than clinical trials for some variables (ECOG-PS, lab and biomarker etc)

ECOG-PS is sometimes missing in real world

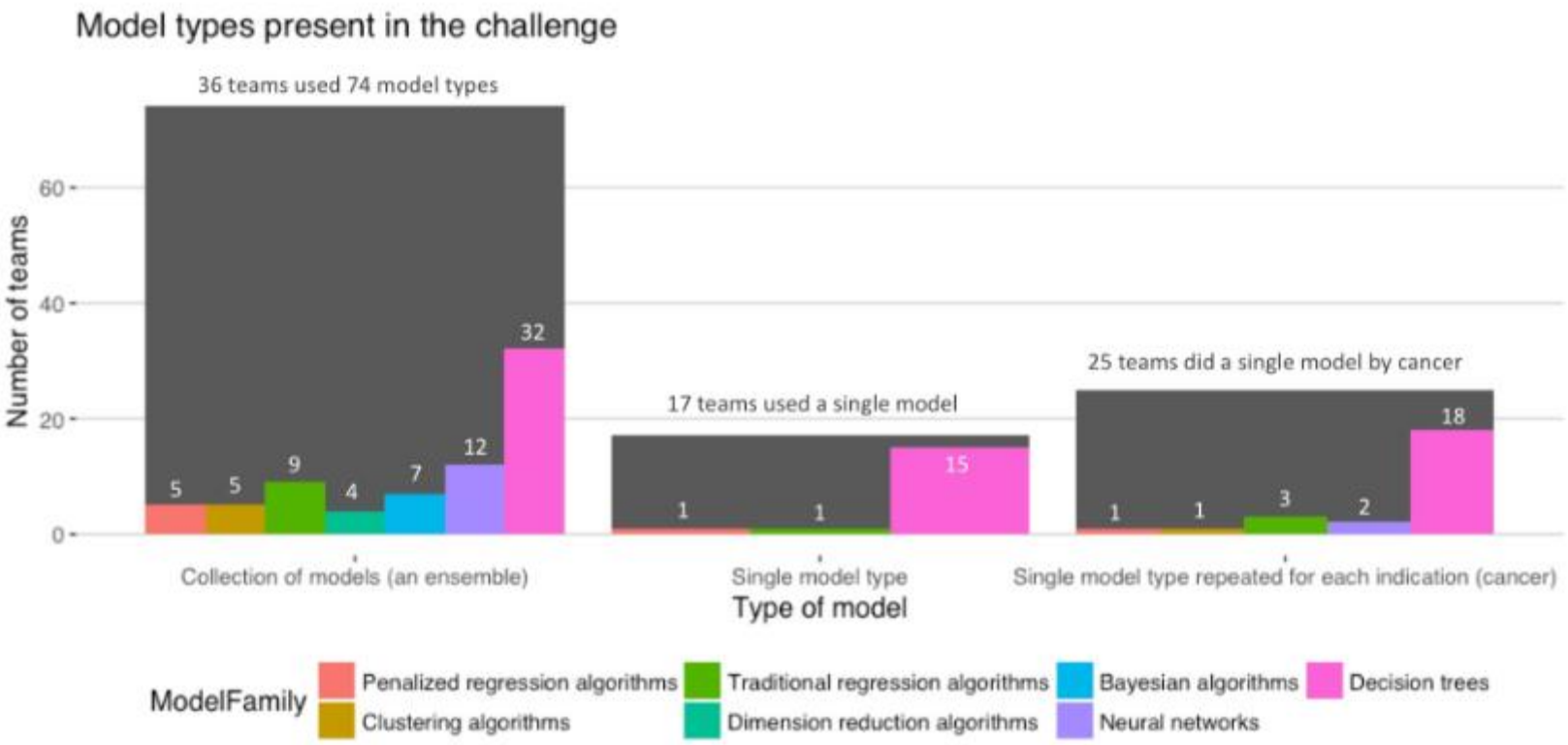


In clinical trials we use specially designed measures to describe a patient's health (e.g. ECOG performance status). In the real world, these measures are captured inconsistently and sometimes missing.

Note: plot for 7000 patients from training set

RAAD Challenge:
Approaches/Models from different teams

Models Used in the Challenge



Decision trees were most often used methods and the winning teams used **gradient boosted decision tree** (XGBoost)

Nomogram approach from team “Predict”

1. The potential prognostic factors were discussed and selected together by statistician and clinician scientists.
2. The statistical models (basically **logistic regression**) were built using 80% “TRAIN” data (select variables by LASSO or relative importance)
3. The models were assessed by 20% “TRAIN” data (with known survival status)
4. The models were discussed with clinician scientists again to check if the point/score assigned for each predictor (prognostic factor) make sense or not, refine the models based on the feedback (add or remove predictors)
5. Build models using all “TRAIN” dataset and predict the probability of survival for “TEST” dataset

Breast Cancer Example

Tumour Type, Points

HER2+, 0
HR+, 39
TNBC, 99
Unknown, 88

CNS metastasis, Points

CNS, 49
None, 0

Liver metastasis, Points

Liver, 45
None, 0

ECOG-PS Evidence, Points

Bad ECOG-PS , 56
None, 0

…

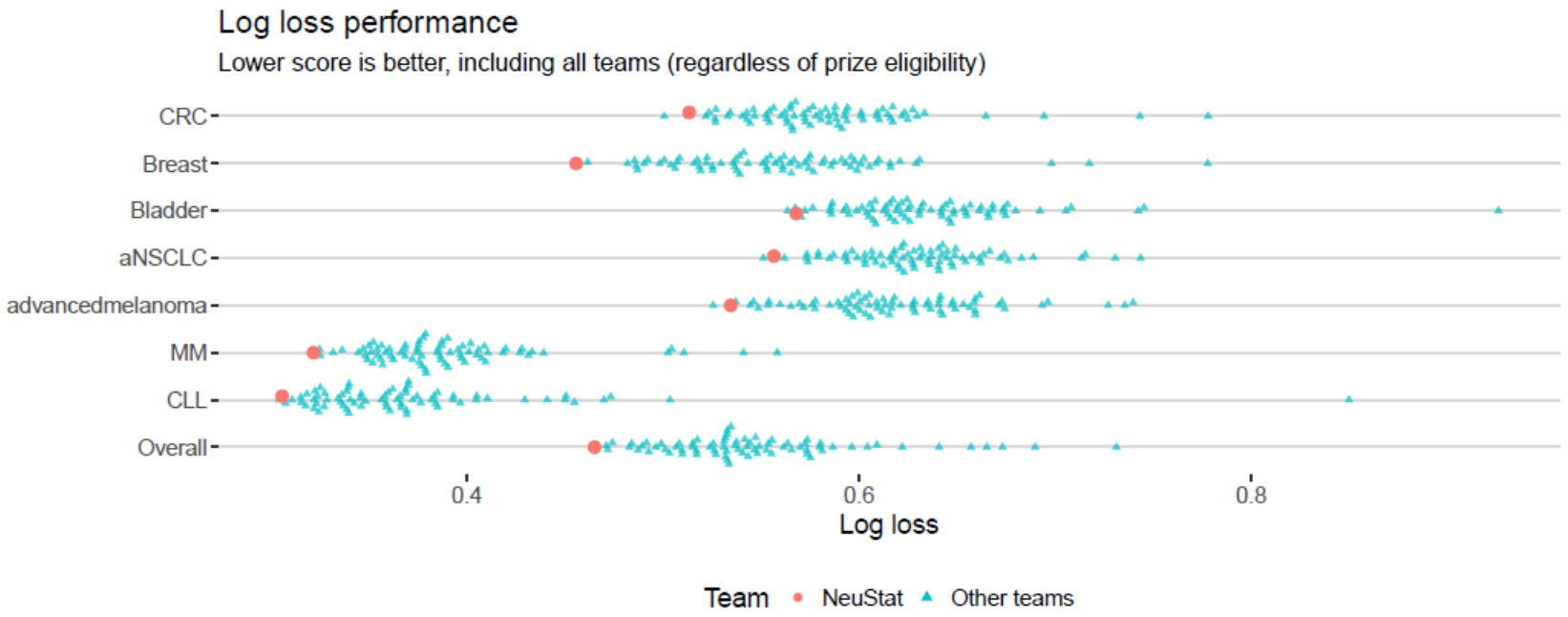
Total Points,	Risk of Death
60,	0.05
95,	0.10
…	
215,	0.60
235,	0.70
260,	0.80
297,	0.90
332,	0.95

XGBoost approach that Team “NeuStat” (and other two top teams) used

- **XGBoost** is short for “Extreme Gradient Boosting”, it is an implementation of gradient boosting decision tree algorithm
- Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction.
- It can handle missing data internally
- Mitigating overtraining with early stopping (or reduce bias by cross-validation and bootstrapping)
- Team NeuStat selected the top 120 features by importance while 2nd and 3rd teams used thousands of features.

RAAD Challenge:
Comparing simple model vs advanced machine learning

Model performance from team NeuStat

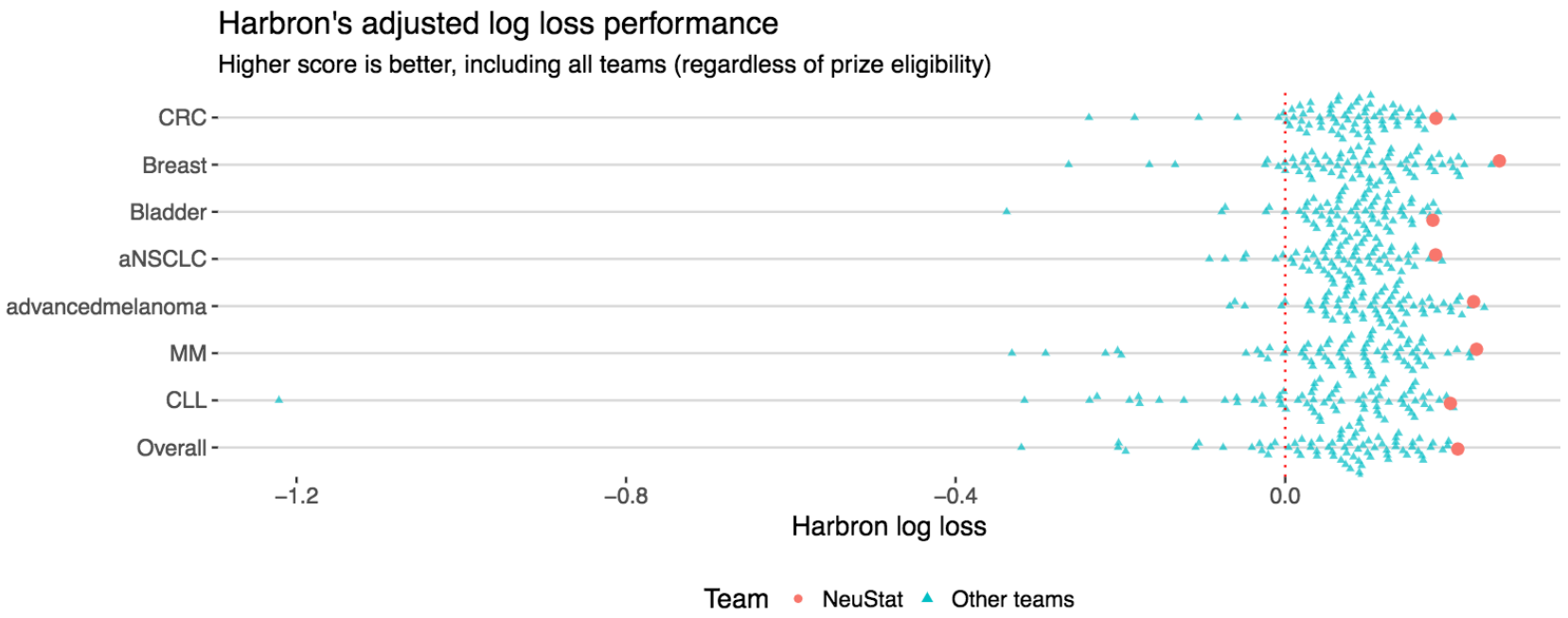


Team NeuStat achieved consistently high accuracy metrics across all cancer types.

The formula for Log Loss for binary classification
(p is predicted probability of death here, y is observed survival status: 0 indicating alive or 1 indicating died):

$$-(y \log(p) + (1 - y) \log(1 - p))$$

Model performance from team NeuStat (cont.)



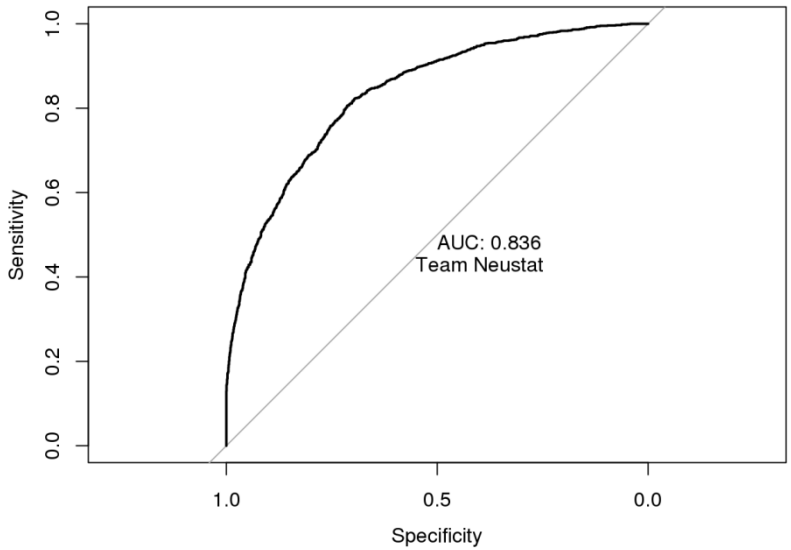
- By using adjusted log loss, you can compare the model performance across different cancer types.
  - The adjusted log-loss puts results from datasets with different response rates onto a common scale
  - A perfect prediction (log-loss=0) obtains a score of 1 and a prediction from the null model obtains a score of 0.
- Team NeuStat's model performed best in breast cancer.

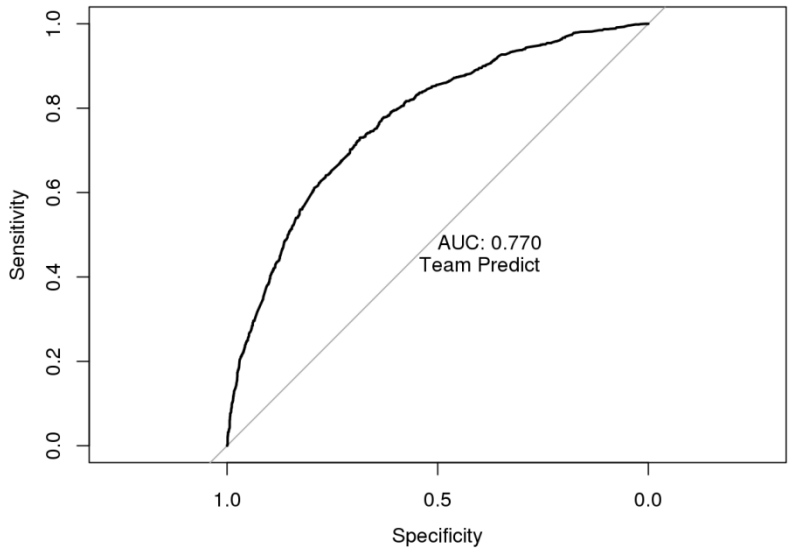
Comparing Machine Learning vs Simple Model

Cancer	This team	Comparison	
		Winner	Diff.
Log Loss			
Overall	0.547	0.465	0.082
Bladder	0.646	0.568	0.078
Breast	0.559	0.456	0.104
CLL	0.369	0.306	0.063
CRC	0.585	0.513	0.071
MM	0.392	0.322	0.070
aNSCLC	0.623	0.557	0.066
advancedmelanoma	0.653	0.535	0.119
Harbron Log Loss			
Overall	0.067	0.209	-0.142
Bladder	0.066	0.179	-0.113
Breast	0.092	0.260	-0.168
CLL	0.036	0.200	-0.165
CRC	0.069	0.183	-0.114
MM	0.066	0.232	-0.166
aNSCLC	0.085	0.183	-0.097
advancedmelanoma	0.057	0.229	-0.171

The performance of XGBoost (decision tree) is better than Nomogram (Logistic Regression) in regarding to LogLoss and adjusted LogLoss score

Comparing Machine Learning vs Simple Model (cont.)





The performance of XGBoost (decision tree) is better than Nomogram (Logistic Regression) in regarding of AUC.

Summary/Learnings

- Advanced Machine Learning outperformed traditional statistical model in this example
  - Better performance in prediction regarding AUC (or LogLoss score)
  - Can use wealthy data (while Lasso etc tried to reduce number of variables in the logistic regression model)
- But machine learning is not as easy as traditional model to be interpreted
- Common challenge, difficult to identify died patients when the mortality rate is low, i.e sensitivity is low. Take CLL as example (only 10% patients died within one year), team NeuStat identified 13% (9/64) died patients when using cut-off of probability as 0.5
- With wealthy real world data is available and the power of machine learning for such data, statisticians would be benefit to know machine learning methods (R packages such as “caret” and “xgboost” available).

Acknowledgement

- Flatron Health
- James Black and the Roche Advanced Analytics Data (RAAD) challenge team
- RAAD challenge winning team NeuStat: Andreas Reichert, Tao Xu and Ying He
- The Roche RAAN (Advanced Analytics Network)

Q&A