Accelerated Failure Time - Xgboost

Avinash Barnwal, Philip Cho and Toby Hocking

December 22, 2019

1 Transformation boosting machines-Review

1.1 Transformation models

$$F_{Y|X=x}(y) = F_Z(h(y|x)) \tag{1}$$

The conditional transformation function h is monotonic in y.
Box and Cox, shift transformation functions based on the decomposition.

$$h(y|x) = h_Y(y) - \beta(x) \tag{2}$$

featuring a baseline transformation function $h_Y : \Xi \to \mathbb{R}$ and a shift term $\beta : \chi \to \mathbb{R}$ have been studied intensively.
The proportional hazards (with $F_Z(z) = 1 - \exp - \exp(z))$) and proportional odds (with $F_Z(z) = expit(z)$) models are the most well-known representatives of this class of shift transformation models (STM, often also referred to as linear or nonlinear transformation models, depending on the functional form of $\beta(x)$).
This is one way of saying that Accelerated failure time is part shift transformation models, where h(y) is log(y) and $\beta(x)$ is same and error is the accelerated failure time can accommodate normal, logistic and extreme distributions.
Structured additive transformation functions that allow interactions between the two arguments y and x of the form $h(y|x) = \sum_{j=1}^{J} h_j(y|x)$ leading to **conditional transformation models**.
STM has β gradient which is different compared to gradient boosting

2 Proposed loss functions

- survival package manual notation, e.g. $z_i = (y_i - f(x_i))/\sigma$, learn model f to max lik of $z_i \sim N(0, 1)$ or logistic or extreme.
- in terms of generalized linear models, e.g. learn model f to max lik of normal $y_i \sim N(\mu_i, \sigma^2)$ with mean parameter $\mu_i = f(x_i)$, or logistic or extreme.

Model

Assume the data follow below model:-

$$\begin{aligned} \log y_i &= x_i^{'}\beta + z_i\sigma \\ \log \hat{y}_i &= x_i^{'}\hat{\beta} \\ \eta &= x_i^{'}\hat{\beta} \\ z_i &= \frac{\log y_i - \eta}{\sigma} \sim f \end{aligned} \tag{3}$$

where y_i is the uncensored response and $\hat{y}_i$ is the predicted value for i-th observation and σ is the standard deviation of the error.

Normal Distribution

$$\begin{aligned} f(z) &= \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \\ f^{'}(z) &= -zf(z) \\ f^{''}(z) &= -f(z) - zf^{'}(z) \\ F(z) &= \phi(z) \\ F_{Y_i}(y_i) &= F(z_i) \end{aligned} \tag{4}$$

Logistic Distribution

$$\begin{aligned} f(z) &= \frac{e^z}{(1+e^z)^2} \\ w &= e^z \\ f^{'}(z) &= f(z)\frac{1-w}{1+w} \\ f^{''}(z) &= f(z)\frac{(w^2-4w+1)}{(1+w)^2} \\ F(z) &= \frac{e^z}{1+e^z} \\ F_{Y_i}(y_i) &= F(z_i) \end{aligned} \tag{5}$$

Extreme Distribution

$$\begin{aligned} w &= e^z \\ f(z) &= we^{-w} \\ F(z) &= 1 - e^{-w} \\ f^{'}(z) &= f(z)(1-w) \\ f^{''}(z) &= (w^2 - 3w + 1)f(z) \\ F_{Y_i}(y_i) &= F(z_i) \end{aligned} \tag{6}$$

3 Uncensored Data

3.1 Loss Function

$$L_i = -\log \text{lik}_i = -\log(f_{Y_i}(y_i|\eta)) \tag{7}$$

where, $f_\eta(\eta)$ is the probability density function(pdf) of η.
Now using change of variable for probability density function(pdf). We can write the below equations.

Normal Distribution

$$\begin{aligned} f_Y(y|\eta) &= f_Z(z)\frac{\partial z}{\partial y} \\ f_Y(y|\eta) &= \frac{\exp^{-}\frac{(\log(y)-\log(\hat{y_i}))^2}{2\sigma^2})}{y\sigma\sqrt{2\pi}} \\ L_i &= -log(\frac{\exp^{-}\frac{(log(y_i)-\log(\hat{y_i}))^2}{2\sigma^2}}{y_i\sigma\sqrt{2\pi}}) \\ L_i &= -log(\frac{\exp^{-}(\frac{(log(y_i)-\eta)^2}{2\sigma^2}}{y_i\sigma\sqrt{2\pi}}) \end{aligned} \tag{8}$$

Logistic Distribution

$$\begin{aligned} f_Y(y|\eta) &= f_Z(z)\frac{\partial z}{\partial y} \\ f_Y(y|\eta) &= \frac{\exp^{(\frac{(\log(y)-\eta))}{\sigma})}}{\sigma * y * (1+\exp^{(\frac{(\log(y)-\eta)}{\sigma})})^2} \\ L_i &= -\log(\frac{\exp^{(\frac{(\log(y)-\eta)}{\sigma})}}{\sigma * y * (1+\exp^{(\frac{(\log(y)-\eta)}{\sigma})})^2}) \\ &= -\log(\frac{\exp^{(\frac{(\log(y)-\eta)}{\sigma})}}{\sigma * y * (1+\exp^{(\frac{(\log(y)-\eta)}{\sigma})})^2}) \end{aligned} \tag{9}$$

Extreme Distribution

$$\begin{aligned} f_Y(y|\eta) &= f_Z(z)\frac{\partial z}{\partial y} \\ f_Y(y|\eta) &= \frac{e^{\frac{\log(y)-\log(\hat{y_i})}{\sigma}}e^{-e^{\frac{\log(y)-\log(\hat{y_i})}{\sigma}}}}{y\sigma} \\ L_i &= -\log\frac{e^{\frac{\log(y_i)-\log(\hat{y_i})}{\sigma}}e^{-e^{\frac{\log(y_i)-\log(\hat{y_i})}{\sigma}}}}{y_i\sigma} \end{aligned} \tag{10}$$

3.2 Negative Gradient

In Gradient Boosting and Xgboost, we need to calculate negative gradient of loss function with respect to real-value prediction which is $\hat{\eta}$. Here i am changing between $\hat{\eta_i}$ to $\hat{\eta}$ to make it general.

$$\begin{aligned} -\frac{\partial L_i}{\partial \hat{\eta}} &= \frac{\partial log(f_{Y_i}(y_i|\eta))}{\partial \eta} \\ &= \frac{1}{f_{Y_i}(y_i|\eta)} * \frac{\partial f_{Y_i}(y_i|\eta)}{\partial \eta} \end{aligned} \tag{11}$$

Using change of variable between z and η. We can write the below in terms of z.

$$\frac{\partial f_Y(y|\eta)}{\partial \eta} = \frac{\partial}{\partial \eta}[f_Z(z)\frac{\partial z}{\partial y}] \tag{12}$$

Using product rule of differentiation, we can split $f_Z(z)$ and $\frac{\partial z}{\partial \eta}$

$$= \frac{\partial f_Z(z)}{\partial \eta}\frac{\partial z}{\partial y} + f_Z(z)\frac{\partial^2 z}{\partial \eta \partial y} \tag{13}$$

$$= \frac{\partial f_Z(z)}{\partial z}\frac{\partial z}{\partial \eta}\frac{\partial z}{\partial y} + f_Z(z)\frac{\partial^2 z}{\partial \eta \partial y} \tag{14}$$

As we know $\frac{\partial f_Z(z)}{\partial z} = -z f_Z(z)$
Now, we will calculate the gradient of z with respect to η as we have $\frac{\partial z^2}{\partial \eta \partial y}$ in the equation 14.

$$\frac{\partial z}{\partial \eta} = \frac{-1}{\sigma} \tag{15}$$

$$\frac{\partial z}{\partial y} = \frac{1}{y\sigma} \tag{16}$$

We also need double differentiation of z with respect to η as we have $\frac{\partial^2 z}{\partial \eta^2}$ in the equation 14.

$$\frac{\partial^2 z}{\partial \eta \partial y} = 0 \tag{17}$$

$$= -\frac{f_Z^{'}(z)}{y\sigma^2} \tag{18}$$

Now going back to original equation of calculating negative gradient of loss function with respect to η. In equation 9, we have calculated the second part of negative gradient of loss function with respect to η which is $\frac{\partial f_{Y_i}(y_i|\eta)}{\partial \eta}$. By replacing the 18th equation in the negative gradient of loss function w.r.t. to η, we get the results as follows:-

$$
\begin{aligned}
-\frac{\partial L_i}{\partial \eta} &= \frac{1}{f_{Y_i}(y_i|\eta)} * \frac{\partial f_{Y_i}(y_i|\eta)}{\partial \eta} \\
&= \frac{1}{f_{Y_i}(y_i|\eta)} * \frac{-f^{'}_{Z_i}(z_i)}{y\sigma^2} \\
&= \frac{-f^{'}_{Z_i}(z_i)}{\sigma f_{Z_i}(z_i|\eta)}
\end{aligned} \tag{19}
$$

3.3 Hessian

Hessian is the second derivative of Loss with respect to η. As we have already calculated the negative gradient, further we need to take one more partial derivative with respect to η of negative gradient with negative sign.

$$
\begin{aligned}
\frac{\partial^2 L_i}{\partial \eta^2} &= \frac{\partial}{\partial \eta}\frac{\partial L_i}{\partial \eta} \\
&= \frac{\partial}{\partial \eta}\frac{f^{'}_{Z_i}(z_i)}{\sigma f_{Z_i}(z_i|\eta)} \\
&= \frac{\partial}{\partial z_i}\frac{f^{'}_{Z_i}(z_i)}{\sigma f_{Z_i}(z_i|\eta)}\frac{\partial z_i}{\partial \eta} \\
&= \frac{\partial}{\partial z_i}\frac{f^{'}_{Z_i}(z_i)}{\sigma f_{Z_i}(z_i|\eta)}\frac{\partial z_i}{\partial \eta} \\
&= -\frac{f_{Z_i}(z_i)f^{''}_{Z_i}(z_i) - [f^{'}_{Z_i}(z_i)]^2}{\sigma^2 f^2_{Z_i}(z_i)}
\end{aligned} \tag{20}
$$

4 Left Censored Data

4.1 Loss Function

$$
L_i = -\log \text{lik}_i = -log(F_{Y_i}(y_i|\eta)) \tag{21}
$$

where, F is the cdf of $Y_i|\eta$.

$$
L_i = -log(F(z_i)) \tag{22}
$$

4.2 Negative Gradient

$$
-\frac{\partial L_i}{\partial \eta} = -\frac{\partial - log(F(z))}{\partial \eta} \tag{23}
$$

$$
-\frac{\partial L_i}{\partial \eta} = \frac{\partial log(F(z))}{\partial \eta} \tag{24}
$$

$$
= \frac{F^{'}(z_i)}{F(z_i)}\frac{\partial z_i}{\partial \eta} \tag{25}
$$

Therefore, combining results of chain rule and pdf, below is the final result for negative gradient,

$$-\frac{\partial L_i}{\partial \eta} = \frac{-f(z_i)}{\sigma F(z_i)} \tag{26}$$

4.3 Hessian

Hessian is the second derivative of Loss with respect to η. As we have already calculated the negative gradient, further we need to take one more partial derivative with respect to η of negative gradient with negative sign.

$$\frac{\partial^2 L_i}{\partial \eta^2} = \frac{\partial}{\partial \eta}\frac{\partial L_i}{\partial \eta} \tag{27}$$

After inputting the values of negative gradient calculated to above equation.

$$\begin{aligned} &= \frac{\partial}{\partial \eta} * \frac{f(z_i)}{\sigma F(z_i)} \\ &= \frac{\partial}{\partial z_i}\frac{f(z_i)}{\sigma F(z_i)} * \frac{\partial z_i}{\partial \eta} \\ &= \frac{F(z_i)f^{'}(z_i) - f^2(z_i)}{\sigma F^2(z_i)} * \frac{-1}{\sigma} \\ &= -\frac{F(z_i)f^{'}(z_i) - f^2(z_i)}{\sigma^2 F^2(z_i)} \end{aligned} \tag{28}$$

5 Right Censored Data

5.1 Loss Function

$$L_i = -\log \text{ lik}_i = -log(1 - F_{Y_i}(y_i|\eta)) \tag{29}$$

where, F is the cdf of $Y_i|\eta$.

$$L_i = -log(1 - F(z_i)) \tag{30}$$

5.2 Negative Gradient

$$-\frac{\partial L_i}{\partial \eta} = -\frac{\partial - log(1 - F(z_i))}{\partial \eta} \tag{31}$$

$$-\frac{\partial L_i}{\partial \eta} = \frac{\partial log(1 - F(z_i))}{\partial \eta} \tag{32}$$

Using similar steps, we have done for left censored.

$$-\frac{\partial L_i}{\partial \eta} = -\frac{F^{'}(z_i)}{1 - F(z_i)}\frac{\partial z_i}{\partial \eta} \tag{33}$$

which is nothing but negative of the negative gradient of left censored data. Therefore,

$$-\frac{\partial L_i}{\partial \eta} = \frac{f(z_i)}{\sigma(1-F(z_i))} \tag{34}$$

5.3 Hessian

$$\frac{\partial^2 L_i}{\partial \eta^2} = \frac{(1-F(z_i))f^{'}(z_i) + f^2(z_i)}{\sigma^2(1-F(z_i))^2} \tag{35}$$

6 Interval Censored Data

6.1 Loss Function

This is combination of left censored and right censored data.

$$L_i = -\log \text{ lik}_i = -log(F_{Y_i^u}(y_i^u|\eta) - F_{Y_i^l}(y_i^l|\eta)) \tag{36}$$

where, F is the cdf of $Y_i|\eta$, y_i^u is the upper limit of time and y_i^l is the lower limit of the time. Above equation is written in terms of ϕ notations as below.

$$L_i = -log(F(z_i^u) - F(z_i^l)) \tag{37}$$

6.2 Negative Gradient

$$-\frac{\partial L_i}{\partial \eta} = -\frac{\partial - log(F(z_i^u) - F(z_i^l))}{\partial \eta} \tag{38}$$

Using chain rule for two variables

$$\frac{\partial L_i}{\partial \eta} = \frac{\partial L_i}{\partial z_i^u}\frac{\partial z_i^u}{\partial \eta} + \frac{\partial L_i}{\partial z_i^l}\frac{\partial z_i^l}{\partial \eta} \tag{39}$$

$$\frac{\partial L_i}{\partial z_i^u} = -\frac{F^{'}(z_i^u)}{F(z_i^u) - F(z_i^l)} \tag{40}$$

$$\frac{\partial L_i}{\partial z_i^l} = \frac{F^{'}(z_i^l)}{F(z_i^u) - F(z_i^l)} \tag{41}$$

$$\frac{\partial L_i}{\partial \eta} = \frac{\partial L_i}{\partial z_i^u}\frac{\partial z_i^u}{\partial \eta} + \frac{\partial L_i}{\partial z_i^l}\frac{\partial z_i^l}{\partial \eta} \tag{42}$$

$$\begin{aligned} -\frac{\partial L_i}{\partial \eta} &= \frac{F^{'}(z_i^u)}{F(z_i^u) - F(z_i^l)}\frac{\partial z_i^u}{\partial \eta} - \frac{F^{'}(z_i^l)}{F(z_i^u) - F(z_i^l)}\frac{\partial z_i^l}{\partial \eta} \\ &= \frac{f(z_i^u)}{F(z_i^u) - F(z_i^l)}\frac{-1}{\sigma} - \frac{f(z_i^l)}{F(z_i^u) - F(z_i^l)}\frac{-1}{\sigma} \\ &= -\frac{f(z_i^u) - f(z_i^l)}{\sigma(F(z_i^u) - F(z_i^l))} \end{aligned} \tag{43}$$

6.3 Hessian

Hessian is the second derivative of Loss with respect to η. As we have already calculated the negative gradient, further we need to take one more partial derivative with respect to η of negative gradient with negative sign.

$$\frac{\partial^2 L_i}{\partial \eta^2} = \frac{\partial}{\partial \eta}\frac{\partial L_i}{\partial \eta} \tag{44}$$

Then we apply the Chain Rule. Since $\partial L_i/\partial \eta$ is now a function of two variables z_i^u and z_i^l, the previous equation is broken down to two components:

$$\frac{\partial^2 L_i}{\partial \eta^2} = \frac{\partial}{\partial \eta}\frac{\partial L_i}{\partial \eta} = \frac{\partial}{\partial z_i^u}\frac{\partial L_i}{\partial \eta}\cdot\frac{\partial z_i^u}{\partial \eta} + \frac{\partial}{\partial z_i^l}\frac{\partial L_i}{\partial \eta}\cdot\frac{\partial z_i^l}{\partial \eta} \tag{45}$$

Let us simplify the first term:

$$\begin{aligned}\frac{\partial}{\partial z_i^u}\frac{\partial L_i}{\partial \eta}\cdot\frac{\partial z_i^u}{\partial \eta} &= \frac{\partial}{\partial z_i^u}\frac{f(z_i^u)-f(z_i^l)}{\sigma\{\phi(z_i^u)-\phi(z_i^l)\}}\frac{\partial z_i^u}{\partial \eta} \\ &= \frac{\{F(z_i^u)-F(z_i^l)\}f^{'}(z_i^u) - f(z_i^u)(f(z_i^u)-f(z_i^l))}{\sigma\{F(z_i^u)-F(z_i^l)\}^2}\frac{-1}{\sigma} \\ &= \frac{-\{F(z_i^u)-F(z_i^l)\}f^{'}(z_i^u) + f^2(z_i^u) - f(z_i^u)f(z_i^l)}{\sigma^2\{F(z_i^u)-F(z_i^l)\}^2}\end{aligned} \tag{46}$$

Similarly, we simplify the second term:

$$\begin{aligned}\frac{\partial}{\partial z_i^l}\frac{\partial L_i}{\partial \eta}\cdot\frac{\partial z_i^l}{\partial \eta} &= \frac{\partial}{\partial z_i^l}\frac{f(z_i^u)-f(z_i^l)}{\sigma\{\phi(z_i^u)-\phi(z_i^l)\}}\frac{\partial z_i^l}{\partial \eta} \\ &= \frac{-\{F(z_i^u)-F(z_i^l)\}f^{'}(z_i^l) + f(z_i^l)(f(z_i^u)-f(z_i^l))}{\sigma\{F(z_i^u)-F(z_i^l)\}^2}\frac{-1}{\sigma} \\ &= \frac{\{F(z_i^u)-F(z_i^l)\}f^{'}(z_i^l) + f^2(z_i^l) - f(z_i^l)f(z_i^u)}{\sigma^2\{F(z_i^u)-F(z_i^l)\}^2}\end{aligned} \tag{47}$$

Combining the terms will give the full expression for $\partial^2 L_i/\partial \eta^2$:

$$\frac{\partial^2 L}{\partial \eta^2} = -\frac{\{F(z_i^u)-F(z_i^l)\}\{f^{'}(z_i^u)-f^{'}(z_i^l)\} - (f(z_i^u)-f(z_i^l))^2}{\sigma^2\{F(z_i^u)-F(z_i^l)\}^2} \tag{48}$$

7 Experiments

7.1 Interval Censored Survival Modeling - Neuroblastoma data-sets

7.1.1 Data

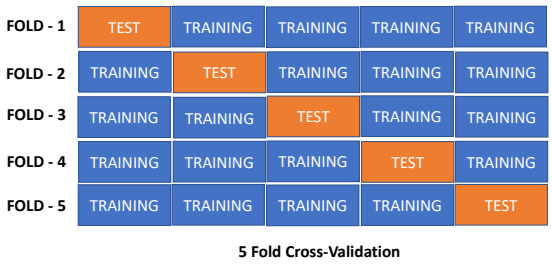
We have used ChIP-seq data generally used for genome-wide peak detection. It has following parts:

- Settings - We have used 5 different settings that lead to 5 different dataset.
- Output - min.log.lambda and max.log.lambda leading to minimum label errors.
- Input - Each attribute is a non-negative integer representing the number DNA sequence reads that has aligned at that particular region of the genome.

5 settings are - ATAC_JV_adipose, CTCF_TDH_ENCODE, H3K27ac_TDH_some, H3K27ac-H3K4me3_TDHAM_BP and H3K36me3_AM_immune.

7.1.2 Cross-Validation

Generally, we have 4-5 folds of the data for each experiment. We have taken each fold as test fold and rest fold is treated as training data. Here, training has been further split into different folds to estimate the hyper-parameters of specific model. This is also called as Nested Cross-Validation. Each test data created for each fold has been used to test the performance of each model. Below is the picture for more details.



5 Fold Cross-Validation

7.1.3 Model

Survival Regression
Multicollinearity leads to inflated coefficients. This gets corrected using uncorrelated variables in the features. Principal components of the features are uncorrelated transformations of the variables. But it leads to another dimension of how many principle components should be used. We have treated

number of principle of components as hyper-parameter. Nested cross-validation mentioned above have been used to tune the hyper-parameter and for each test-fold accuracy is calculated based on equation below :-

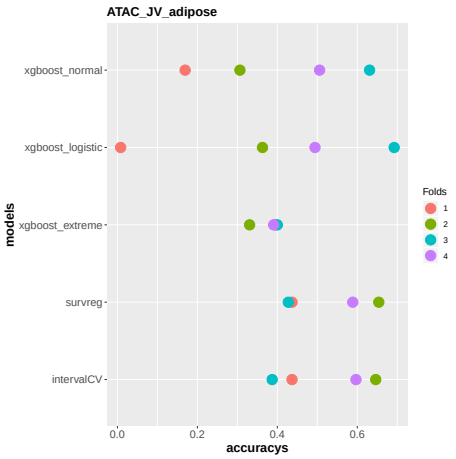
$$accuracy = \sum_{i=1}^{n} I_{y.lower_i < \hat{y_i} < y.upper_i} \tag{49}$$

In the above, we have $y.lower_i$ and $y.upper_i$ are the actual i^{th} response data, $\hat{y_i}$ is the predicted response and n is the number of the observation in the data.

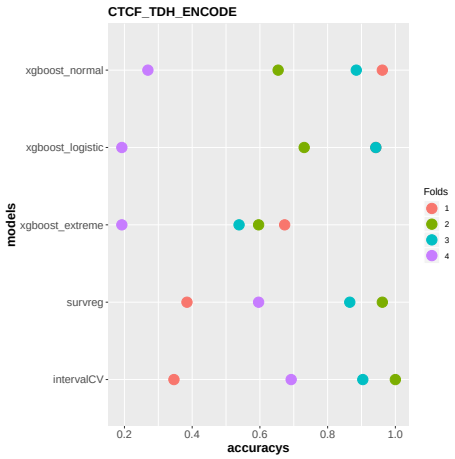
Interval Penalty Learning
It has in-built hyper parameter, we need to pass training data and calculate the accuracy metrics for the test data.

Xgboost
Cox-Ph model in Xgboost doesn't support interval censored data. Accelerated Failure Time in Xgboost supports interval censored data. We have tuned all the hyper-parameters using optuna framework including distribution - Normal, Logistic and Extreme and Sigma.

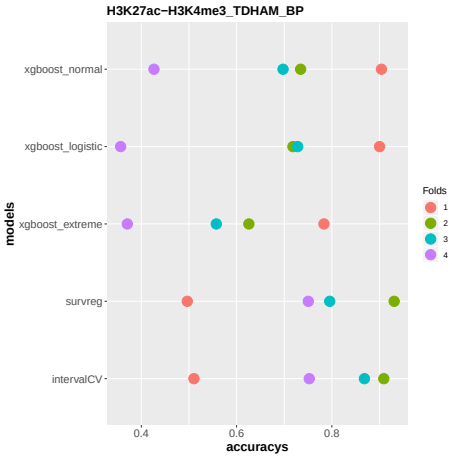
Results
For xgboost, we have kept $\sigma = 1$ for normal and logistic distributions and $\sigma = 10$ for extreme distribution. Extreme distribution doesn't converge with $\sigma = 1$. Following are the accuracy on test data across the folds for 5 data sets mentioned above:-



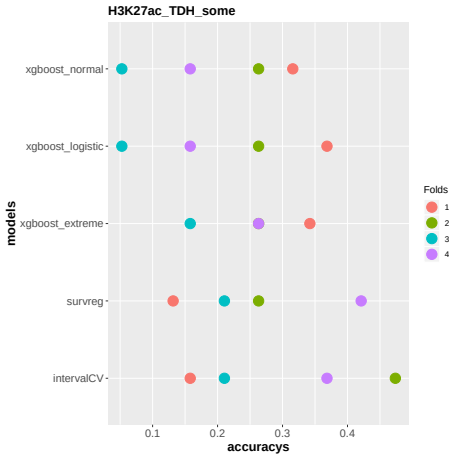
(a) ATAC_JV_adipose - Accuracy



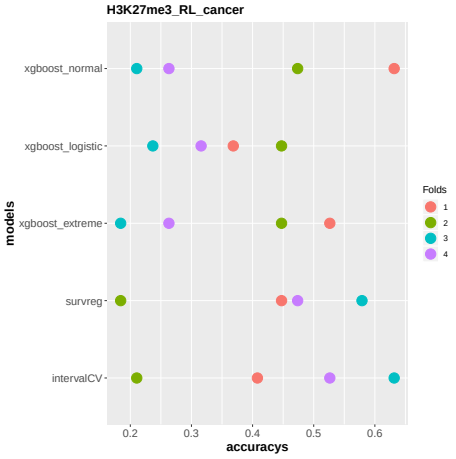
(b) CTCF_TDH_ENCODE - Accuracy



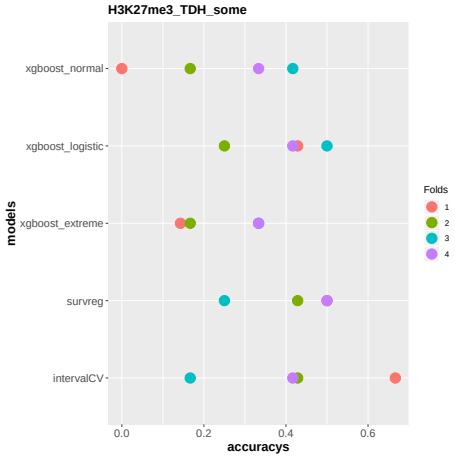
(c) H3K27ac-H3K4me3_TDHAM_BP - Accuracy



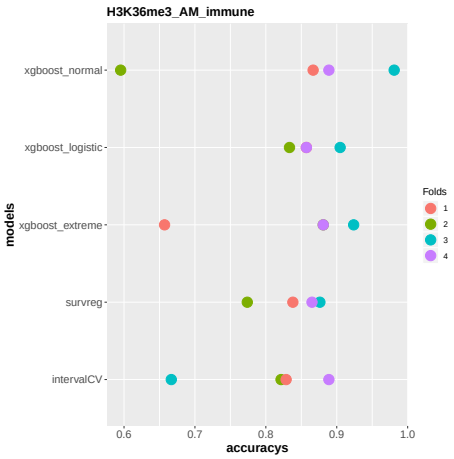
(d) H3K27ac_TDH_some - Accuracy



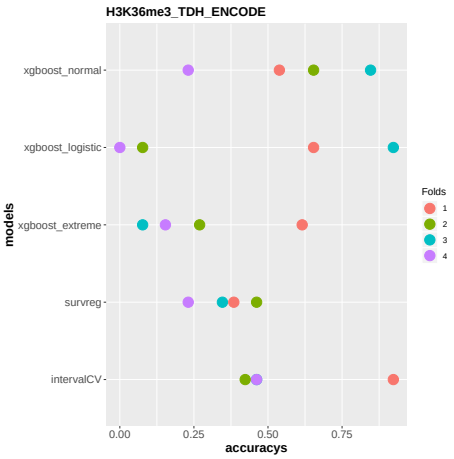
(e) H3K27me3_RL_cancer - Accuracy



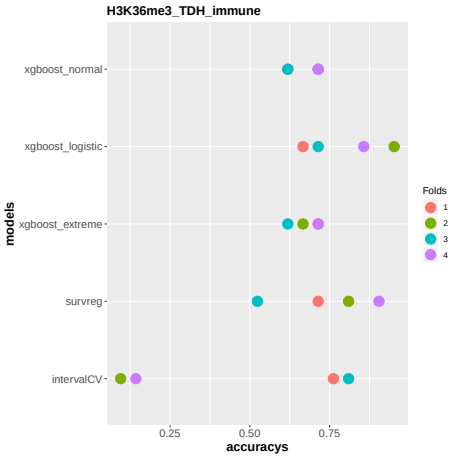
(f) H3K27me3_TDH_some - Accuracy



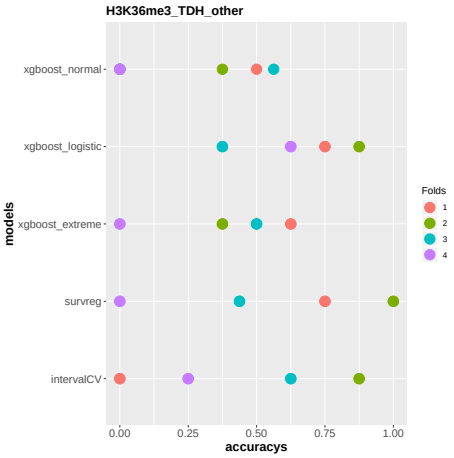
(a) H3K36me3_AM_immune - Accuracy



(b) H3K36me3_TDH_ENCODE - Accuracy

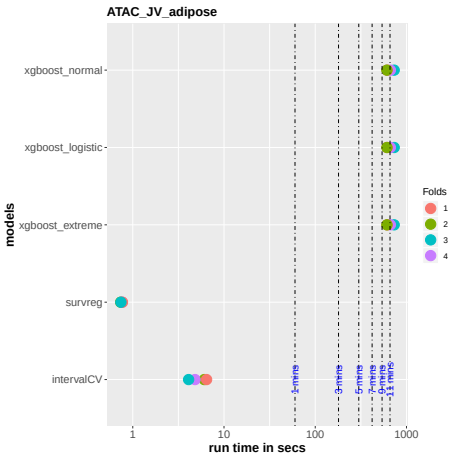


(c) H3K36me3_TDH_immune - Accuracy

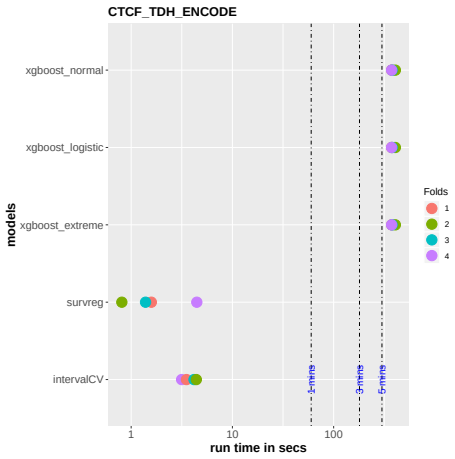


(d) H3K36me3_TDH_immune - Accuracy

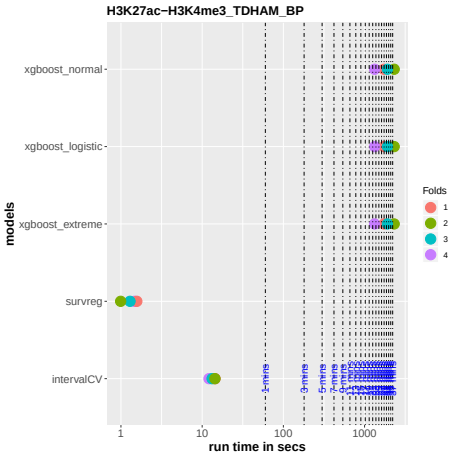
We have generated run time plot for each model. Following are the run time plots for each model with specific distribution for each fold across the datasets:-



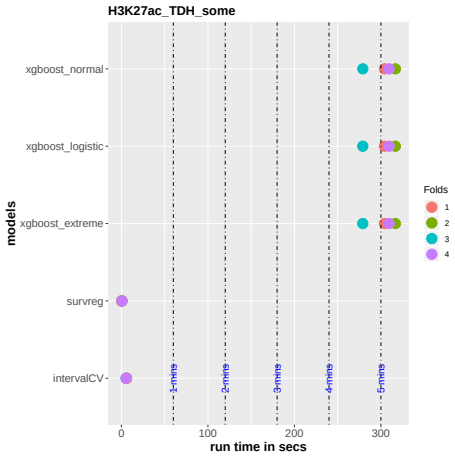
(a) ATAC_JV_adipose - Run Time



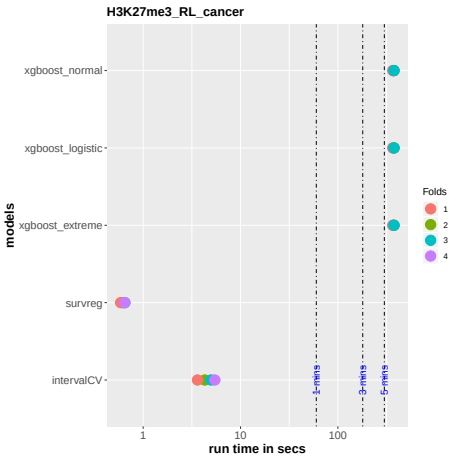
(b) CTCF_TDH_ENCODE - Run Time



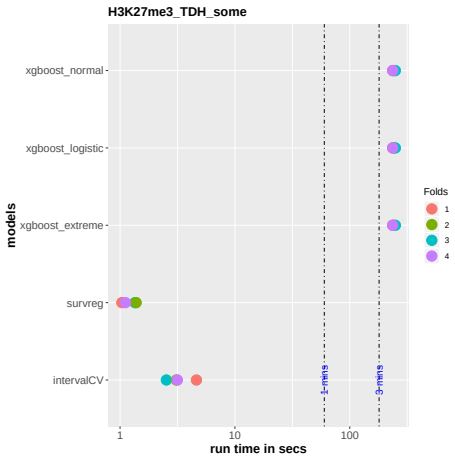
(c) H3K27ac-H3K4me3_TDHAM_BP - Run Time



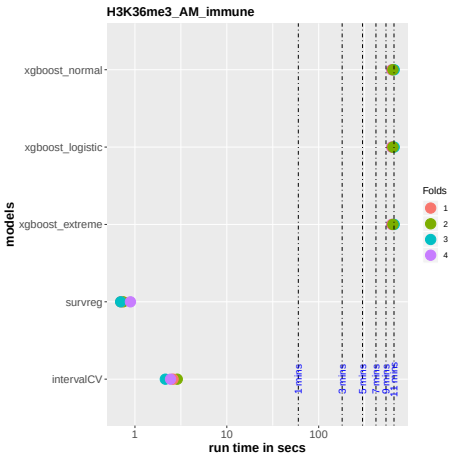
(d) H3K27ac_TDH_some - Run Time



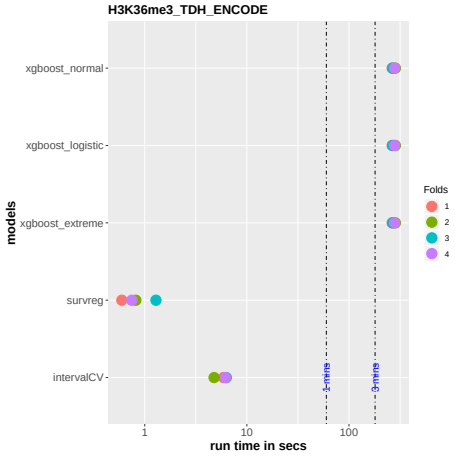
(e) H3K27me3_RL_cancer - Run Time



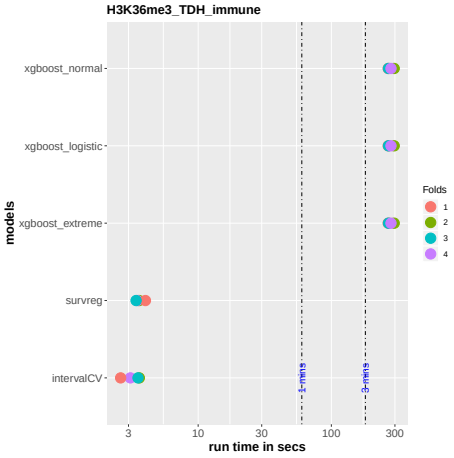
(f) H3K27me3_TDH_some - Run Time



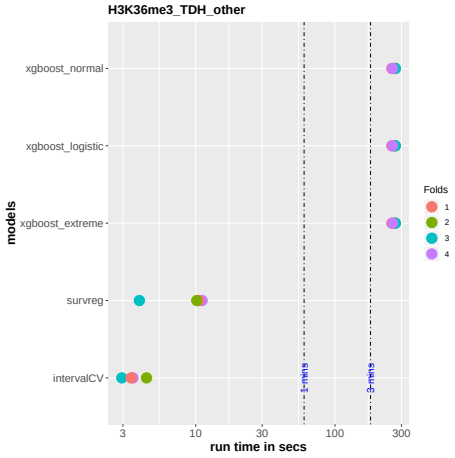
(a) H3K36me3_AM_immune - Run Time



(b) H3K36me3_TDH_ENCODE - Run Time



(c) H3K36me3_TDH_immune - Run Time



(d) H3K36me3_TDH_immune - Run Time

Extra Notes -
H3K36me3_TDH_immune - Normal - $\sigma = 10$
H3K36me3_TDH_other - Normal - $\sigma = 10$

References