Neuron Aging State Prediction Using the Expression Profiles of X-chromosome Genes

Doudou Yu
Data Science Initiative, Brown University
Github: https://github.com/ddfishbean/data1030_project.git

I. Introduction

Aging is one of the biggest risk factors for diseases such as neurodegenerative disease[1]. As the aged population grows, the research on the biology of aging has gained tremendously increased resources. However, it is not an easy task to evaluate an anti-aging intervention, as the golden standard for testing the effectiveness is lifespan extension[2]. Most of the time, it takes years to do lifespan experiments on mice. One way to tackle this problem is to leverage machine learning tools to learn the gene expression profiles of young and aged cells and predict whether a certain intervention can rejuvenate the cells based on the gene expression profile. In this *classification* problem, the *target variable* is 'young' or 'aged'.

This project uses gene expression profiles from neuronal cells in the mouse hypothalamus -- a brain region that controls sleep patterns, energy expenditure, and homeostasis[3]. During aging, dysregulation of the hypothalamus contributes to common age-related phenotypes such as changes in body weight. Neuronal cells are of particular interest because 1) they cannot regenerate after damage or degeneration, 2) there are barely any effective treatments for neurodegenerative diseases, and 3) we can directly generate neurons using easily accessible skin cells in a petri dish to evaluate neuronal gene expression profile[4]. For the *features*, I selected X-chromosome genes because males and females age differently and one distinct genetics is that females have two X chromosomes, while males have one. To balance the dosage, one of the X chromosomes in females is randomly silenced during embryonic development. However, during aging, this X chromosome silencing could be disrupted and contribute to aging, which could be aging markers to predict aging status[6].

The dataset is generated by the Webb laboratory at Brown[5]. It is a single nuclei RNA sequencing dataset that has 25,002 cells (*data points*) with each cell as a row and 283 columns. The dataset has eight *female* mouse hypothalami. Note that for this project, the goal is to *predict the aging state of cells from these eight animals*, but not previously unseen animals, therefore no group structure is included. There are three different kinds of features. First, the *categorical* column -- **'tree.ident'**, stores neuronal subtypes such as Sst/Npyl. Second, 278 columns with *continuous* features, each with the expression (*numerical*, the unit is the number of reads) of an individual X-chromosome gene. Third, engineered features that combine the other features. **'x_sum'** is a *numerical* column that shows the *aggregated expression* of X-chromosome genes, **'sum'** shows the expression of all genes from one cell, and **'x_prop'** is calculated as x_sum/sum representing the percentage of X genes expression. Note that I filtered out X-chromosome genes that have a low expression (the mean expression < 0.1 reads) or expressed in less than 3,000 cells.

II. Exploratory Data Analysis (EDA)

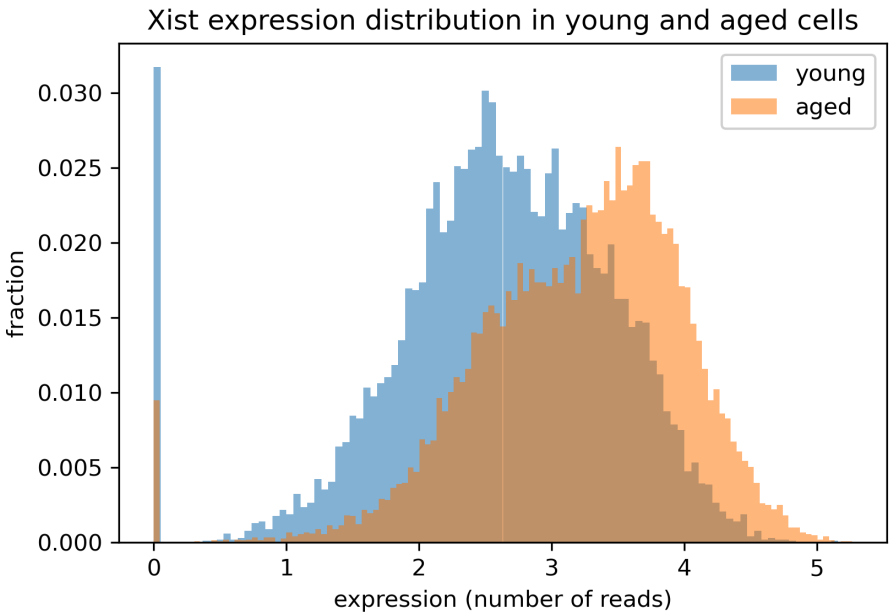

Figure 1. The histogram distribution of the expression of *Xist* grouped by young (blue) and aged (yellow) labels. Both distributions follow the zero-inflated negative binomial distribution. Interestingly, aged cells have higher mean expression than young cells, and the two distributions overlap for around 70% of the total area. Besides, both distributions have similar variance but there are much more young cells with zero expression than the aged cells.

Xist is a non-coding RNA that silences the inactivated X-chromosome. Overall, aged cells have higher *Xist* expression, which could be a compensatory mechanism for keeping the silencing of the inactivated X-chromosome. Besides, the two distributions overlap a lot, which could result from: during normal aging, 1) most cells have increased *Xist* expression, 2) a subset of cells have a very dramatic increase of *Xist* expression, similar to aging pioneer cells.

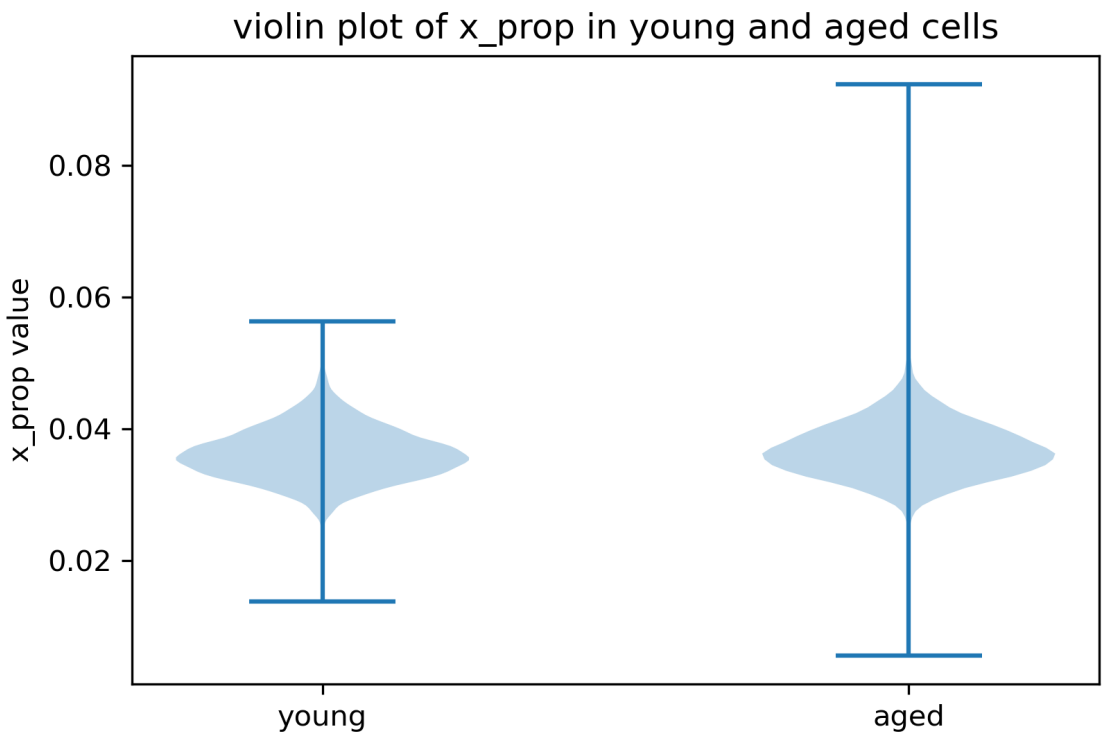

Figure 2. Violin plot shows the x_prop value in young and aged cells. Larger x_prop means that the overall gene expression is skewed towards the X chromosome, indicating that there might be loss of the X chromosome silencing. The young and aged cells have similar distribution and mean, but the aged cells have a larger variance.

Figure 1 showed that during aging, *Xist* upregulation may help control the loss of silencing in aging. Figure 2 showed that there are similar x_prop values in both ages, indicating that most of the aged cells do not lose X-chromosome silencing. Interestingly, a small proportion of the aged cells have very extreme x_prop values, indicating the possibility that a small number of cells age at first and then drive the aging process forward.

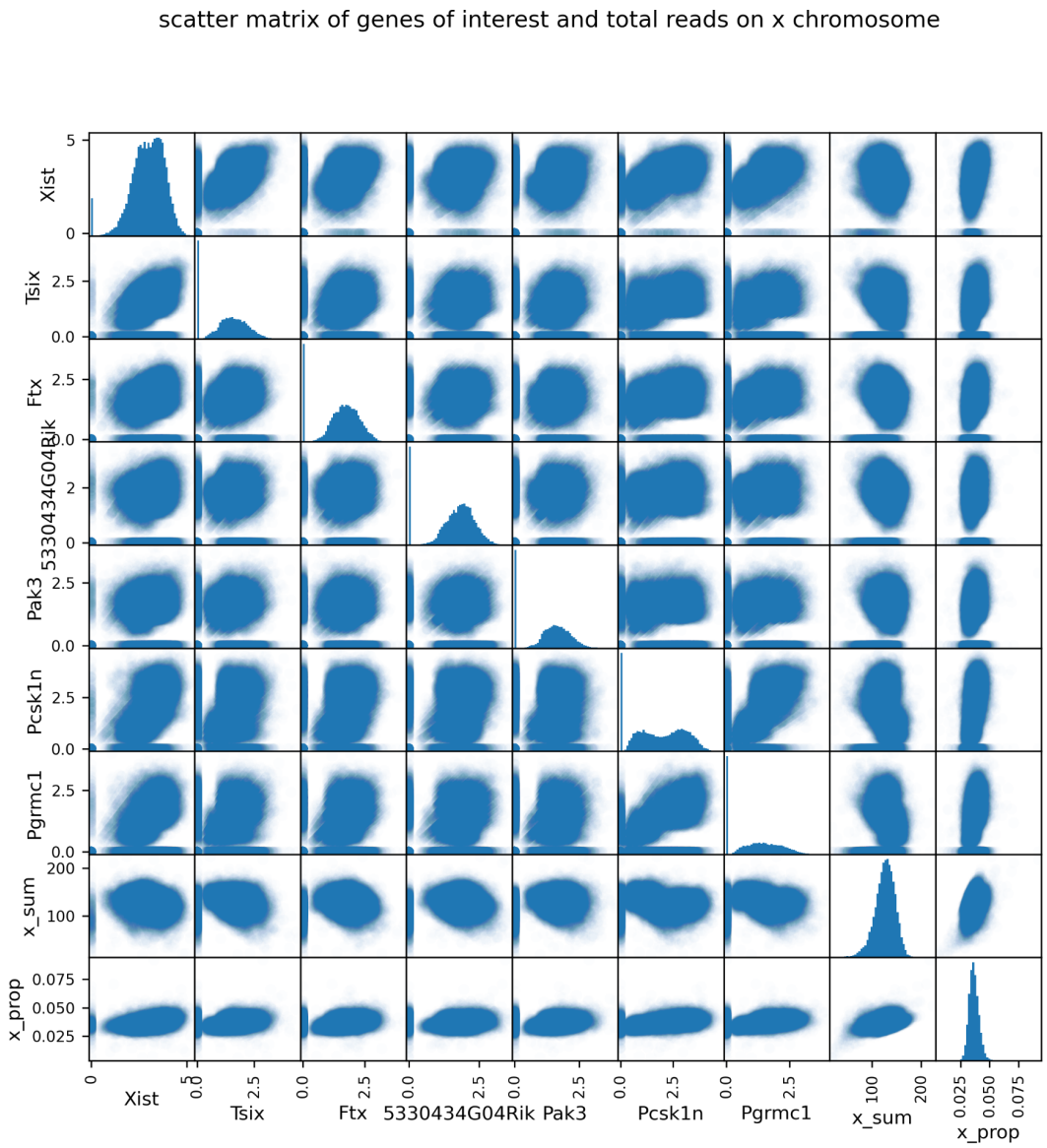

Figure 3. The scatter matrix showing the relationship between genes of interest, and gene expression vs x_prop or x_sum. Most of the genes follow the zero-inflated negative binomial distribution, but the Pcsk1n looks like an overlap of two binomial distributions. This could result from batch effects or differences between different samples.

Several genes have distinct expressions in young versus aged cells. The scatter matrix is then plotted using these genes together with x_sum and x_prop. *Tsix*, which stops *Xist* from silencing other genes[6], has a positive correlation with *Xist*. This raises an interesting question: is *Tsix* upregulation a way to counteract *Xist* upregulation? Besides, this plot shows some feature interactions, indicating the difficulties in selecting features solely based on dependency tests.

III. Methods

1) Data Splitting and Preprocessing

The data is independently identically distributed with no time-series property. Because although it has group structure (eight mice), the goal is not to predict previously unseen animals, but to predict whether the cell *in the eight animals* is “young” (0) or “aged” (1) (target variable). I first split 20% observations into the *testing set*, and 80% for further five fold cross-validation to account for the variability in different random states and relatively small data set. For each cross-validation, 80% of the observations is for the *training set*, and 20% for *validation*. As a result, train-validation-testing is 64-16-40. To avoid data leakage, I used the GridSearchCV method, which first fit and transformed the training set, and then transformed the validation and testing sets.

For preprocessing the features, the continuous features such as gene expression is very zero-inflated and not clearly bounded, so StandardScaler should be a good fit. For the unordinary categorical feature, I used OneHotEncoder. In total, there are 282 features (281 continuous + 1 categorical with 34 categories). To best mimic future model deployment, I tried ten random states to check how much the random split affects test results and reproducibility. I also avoided using the information in the validation or test set when deploying the model.

2) Model Selection and the Final Model Formulation

With ten different random states and previously mentioned preprocessing strategy, I developed an ML pipeline, and tried eight models with corresponding hyperparameters (except for Naive Bayes) shown below. I tuned the hyperparameters with GridSearchCV, extracted the best parameters, and evaluated the models’s performance based on accuracy score as well as f scores with beta=0.5 (put more emphasis on precision since it is expensive to perform the anti-aging interventions) on the test sets.

Tabel 1. Parameters used for tuning models

Model	Parameters
L1 (Lasso)	**C**: 0.01, 0.05, **0.1**, 0.5, 1, 5, 10
L2 (Ridge)	**C**: 0.01, 0.05, **0.1**, 0.5, 1, 5, 10
ElasticNet	**C**: 0.05, **0.1**, 0.5, 1, 3; **l1_ratio**: 0.2, 0.35, **0.5**, 0.65, 0.8
Random Forest	**max_features**: 25, 50, **75**, 100, 200, None; **max_depth**: 10, 20, **30**, 50, 100, None; **min_samples_split**: 2, 5, **10**, 20
SVC	**gamma**: 1e-4, 1e-3, **1e-2**, 1e-1; **C**: np.logspace(-1, 1, 5) **(3.162)**
XGBoost	**max_depth**: 1, 3, **5**, 10, 20, 30, 100; 1, 2, 3, 4, **5**, 8, 10, 15, 20 (finer)
KNN	**n_neighbors**: 30, 100, 200, 300; **weights**: uniform, distance

SVC: Support Vector Machine Classifier with rbf kernel,
KNN: K Nearest Neighbor
The best parameters are bolded or indicated next to the grid.

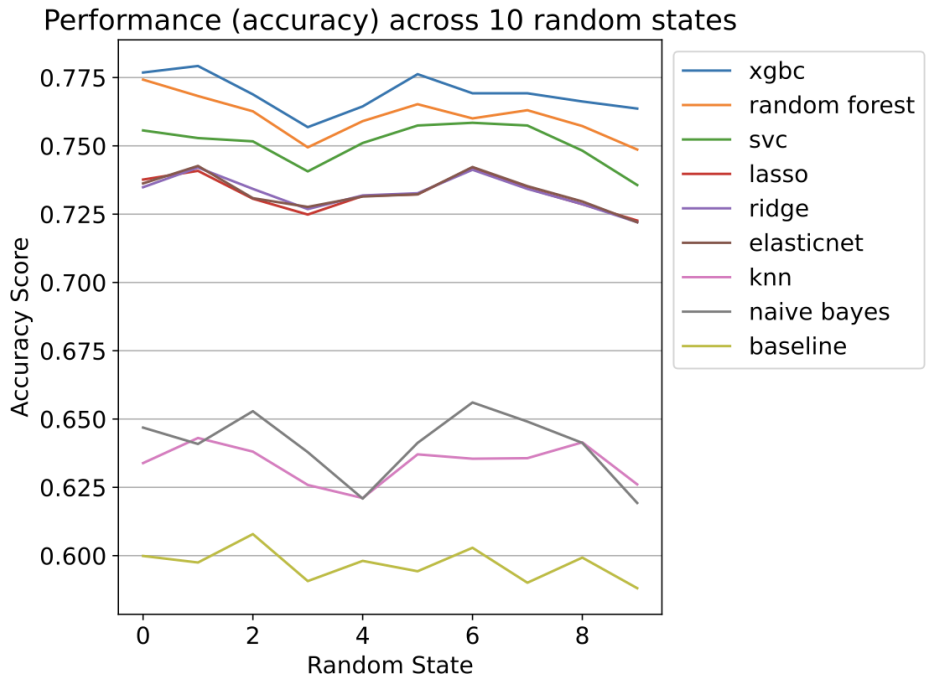

Figure 4. The line chart showing the model performance (accuracy) on the same test sets across 10 random states. XGBoost classifier (xgbc) performs the best and the knn performs the worst, but all of the models have better predictions than the baseline. Compared with the linear models, the other non-deterministic models have more uncertainties from splitting.

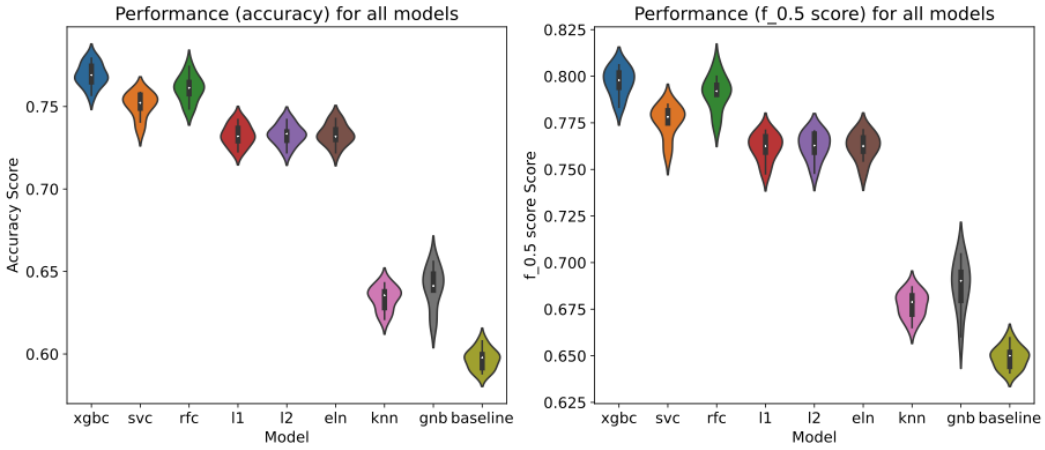

Figure 5. The violin plots showing the model test set performance (accuracy -- left and f_0.5 -- right score) across 10 random states. Xgbc outperforms the others.

Based on Figure 4-5, I chose xgbc for further analysis. To get better parameters, I tuned the xbgc with finer parameters on Table 1 using the early_stop function. Finally, max_depth = 5 was chosen because it has the highest test scores for most of the random states (notebook4). The xgbc model was retrained on new splits across 50 different random states. The

train-validation-testing is still 64-16-20 because the cross validation is needed for the 'early_stop'.

IV. Results of the Best Model (XGBoost)

1) Model Evaluation

For 50 random states, the mean *baseline* accuracy is 0.596 with 0.007 as standard deviation (std), while the mean trained *model* accuracy is 0.778 with 0.006 as std. The mean of the trained model is 0.182 above the mean baseline, which is about 26 std above the baseline. In reverse, the mean baseline accuracy is 30 std below the mean trained model accuracy.

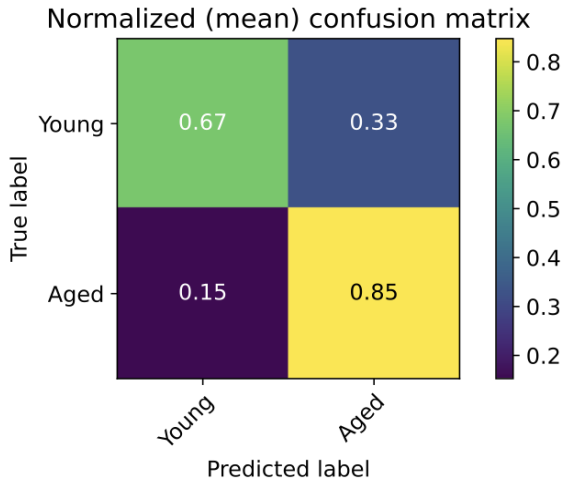

Figure 6. The normalized confusion matrix of the mean value over 50 random states.

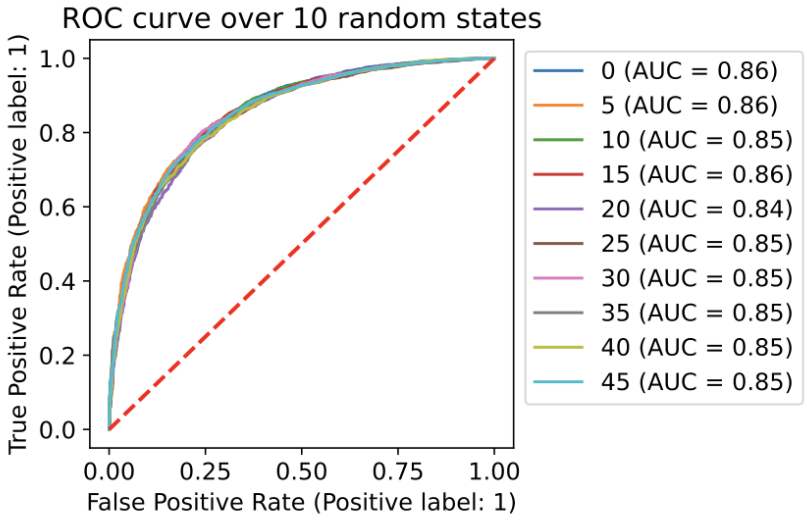

Figure 7. The ROC curve from 10 random states.

Based on the confusion matrix and ROC curve generated using the results from 50 different random states , the model is doing a relatively decent job. The model is better at predicting true positives than the true negatives. This could be due to the fact that the observations have slightly more aged cells than young cells.

2) Model Interpretation - Global Features

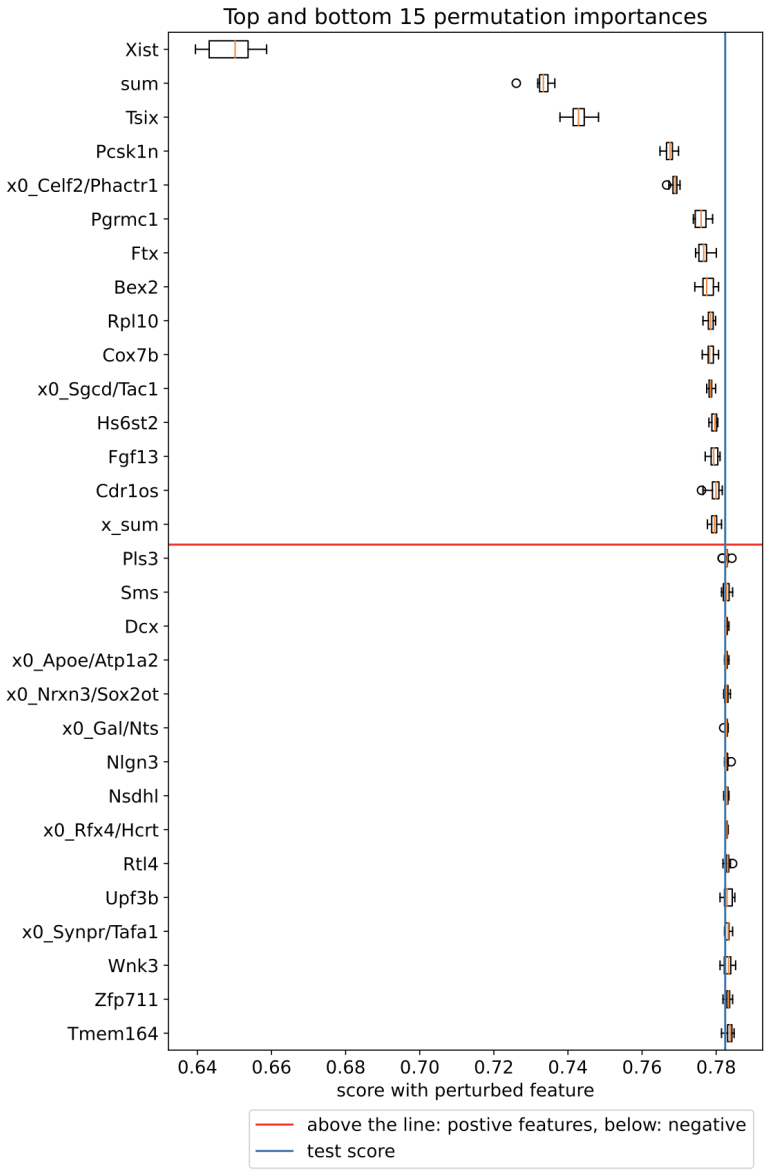

Figure 8. Top and bottom 15 most important features (evaluated on the test set from one random state). The top 15 features are the same over time, while the bottom 15 change a lot.

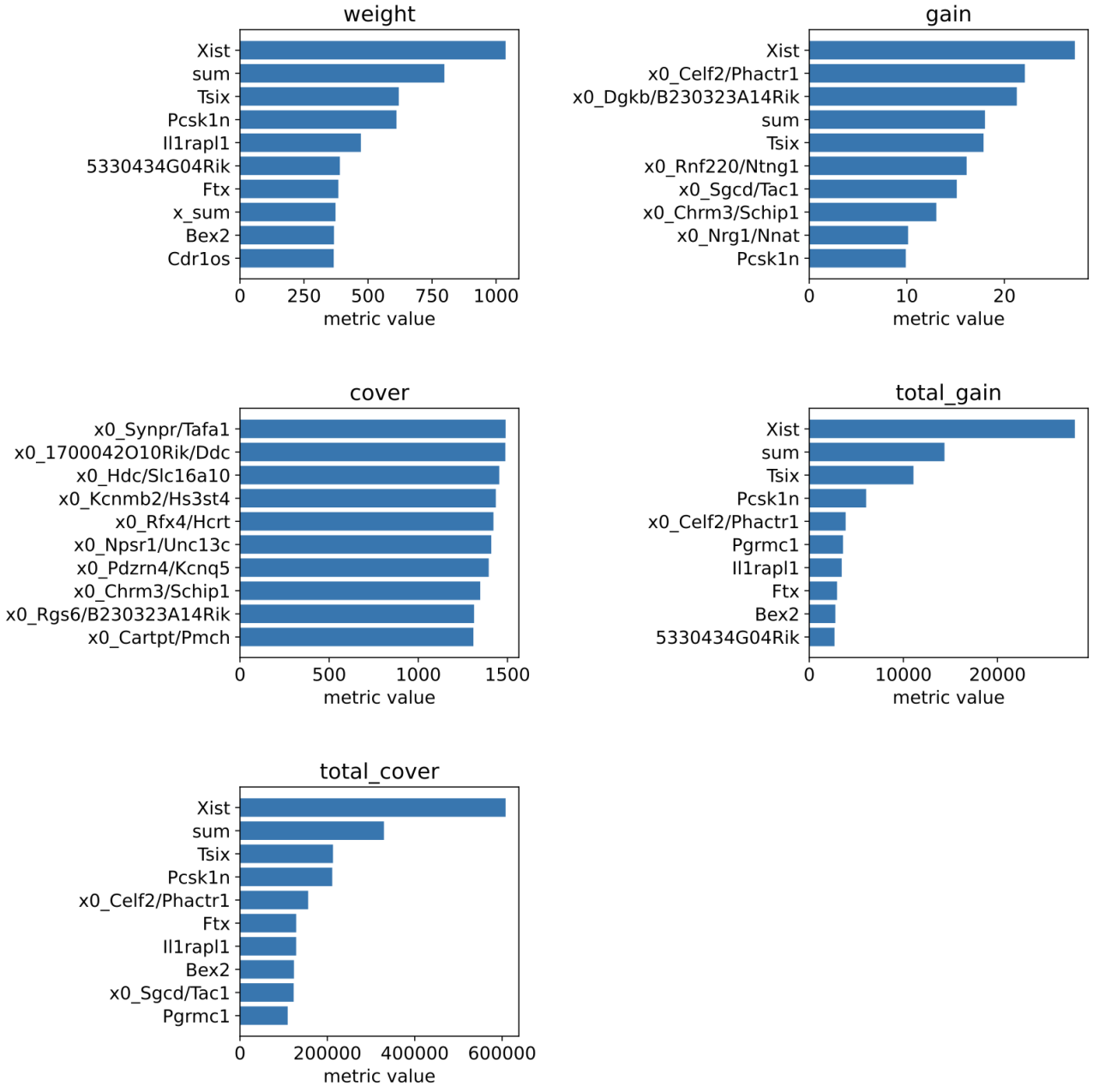

Figure 9. Top 10 most important features for five XGBoost metrics.

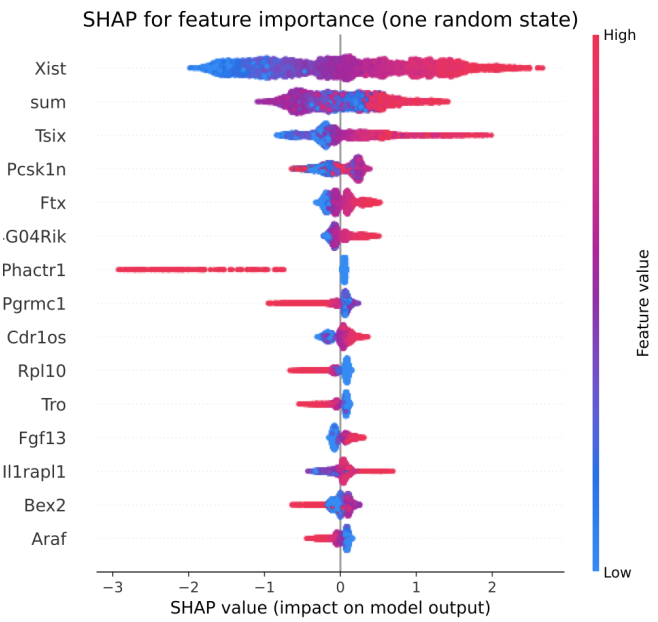

Figure 10. Top 15 most important features using SHAP summary plot.

Based on the three different feature importance, the *most* important features are *Xist*, *Tsix*, *Pcsk1n*, *Ftx*. This echoes what we found during EDA and genomic analysis[5] -- *Xist* and *Tsix* are the most important differentially expressed genes in female aging. The *least* important features are mostly cell types, indicating that there might be some cell-type specific transcriptomic signatures, which could not be generalized to predict the age of other cell-types. Surprisingly, the *sum* of the total gene expression is among the top three important features, meaning that during aging there is global change of the gene expression, which could partially explain why *Xist*, the chromosome silencing master gene, overexpressed in aging.

3) Model Interpretation - Local Features

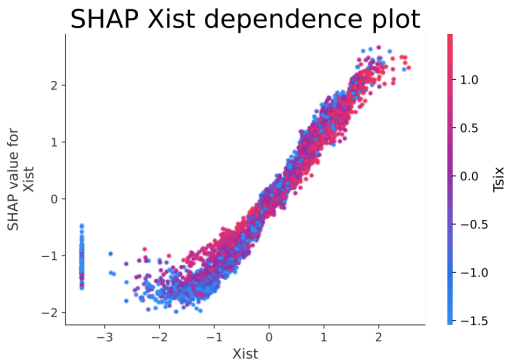

Figure 11. *Xist* dependence plot showing the feature interaction between *Xist* and *Tsix* (similar to what shown in Figure 3).

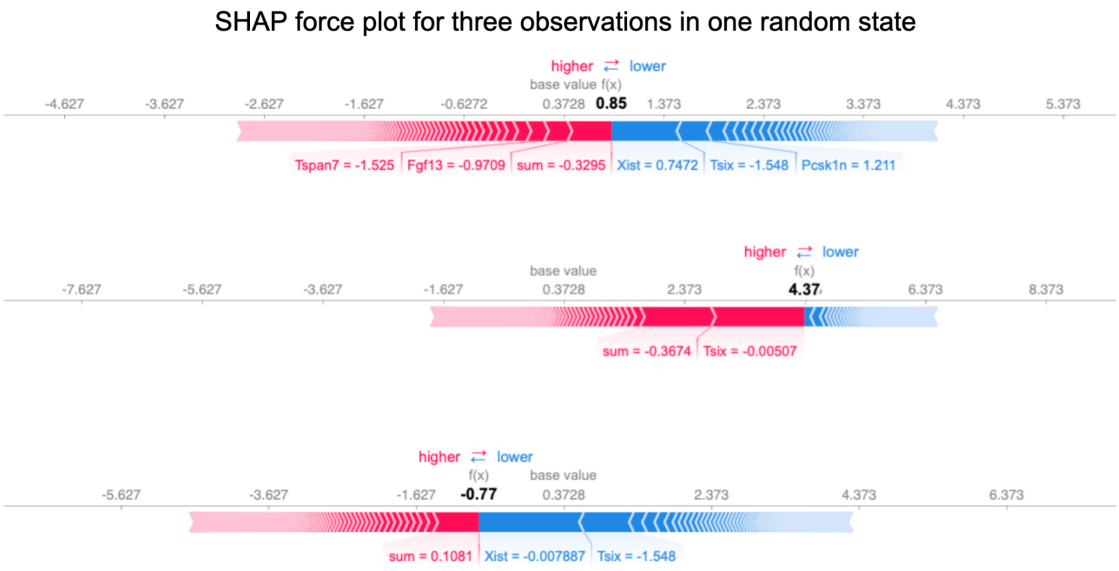

Figure 12. SHAP force plot showing which features had the most impact on the model prediction for three observations (all have right predictions).

The local feature analysis using SHAP values showed that *Xist* and *Tsix* (the top global features) have great impact on the model predictions, and these two features are interacting with each other, which is similar to what the EDA (Figure 3) had found. For the first observation in Figure 12, the model score (0.85) is close to the base value though it is classified right (aged). Unexpectedly, *Xist* and *Tsix* are playing negative roles in this specific aged cell prediction. This reveals the problem that some cells in aged individuals could be resilient to aging, which makes the classification problem tough.

V. Outlook

Though the most and least important features used in the XGBoost tree model correspond to experimental evidence, to increase the interpretability, we can try XGBoost linear models with regularization. To increase the model accuracy, autosomal genes could be included and booster parameters like gamma could be further tuned, together with gpu implementation. The weak spot is the splitting strategy. The goal here is to classify cells in the animals investigated. But for broader applications, to predict cells in *previously unseen animals*, the group-based-splitting should be used, which may improve the model. Grouping should also take different batches and experimental tools into account[7]. Additional techniques could be neural networks but compromised with lower interpretability. Given the low signal to noise ratio of the single cell data, the model will benefit greatly from extra data, especially extra animals, with similar design and strictly controlled variables. With more age time points, regression could be applied to predict a specific age of a cell, rather than binary young or aged.

VI. References:

1. Partridge, L., Deelen, J. & Slagboom, P. E. Facing up to the global challenges of ageing.

Nature 561, 45–56 (2018).
2. Horvath, Steve, and Kenneth Raj. "DNA methylation-based biomarkers and the epigenetic clock theory of ageing." Nature Reviews Genetics 19.6 (2018): 371-384.
3. Ishii, M. & Iadecola, C. Metabolic and Non-Cognitive Manifestations of Alzheimer's Disease: The Hypothalamus as Both Culprit and Target of Pathology. *Cell Metab* 22, 761–776 (2015).
4. Pang, Z. P., Yang, N., Vierbuchen, T., Ostermeier, A., Fuentes, D. R., Yang, T. Q., Citri, A., Sebastiano, V., Marro, S., Südhof, T. C. & Wernig, M. Induction of human neuronal cells by defined transcription factors. *Nature* 476, 220–223 (2011).
5. Hajdarovic, Kaitlyn H., et al. "Single cell analysis of the aging hypothalamus." *bioRxiv* (2021).
6. Stavropoulos, Nicholas, Naifung Lu, and Jeannie T. Lee. "A functional role for Tsix transcription in blocking Xist RNA accumulation but not in X-chromosome choice." Proceedings of the National Academy of Sciences 98.18 (2001): 10232-10237.
7. Whalen, Sean, et al. "Navigating the pitfalls of applying machine learning in genomics." *Nature Reviews Genetics* (2021): 1-13.