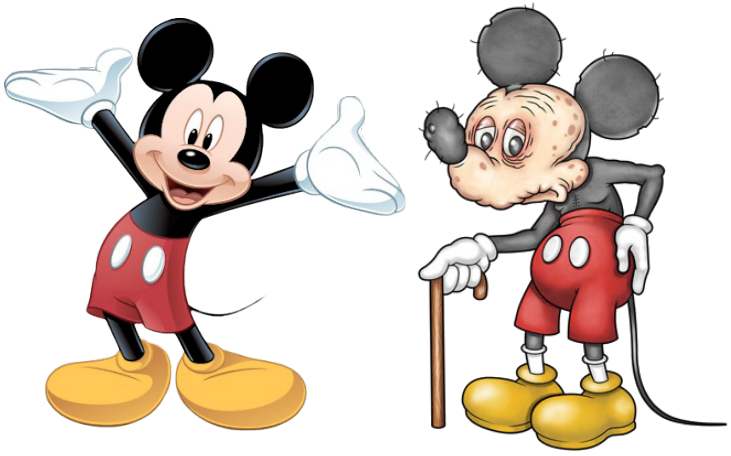
Predicting Young or Aged Cells from the X-chromosome Gene Expression Profile

Doudou Yu
Data Science Initiative
Brown University
2021.12.10

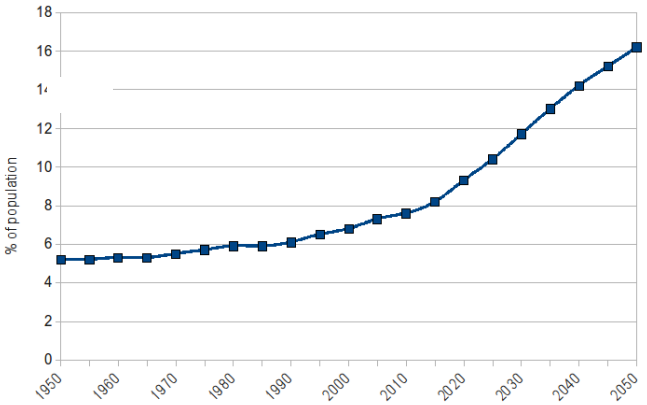


https://github.com/ddfishbean/data1030_project

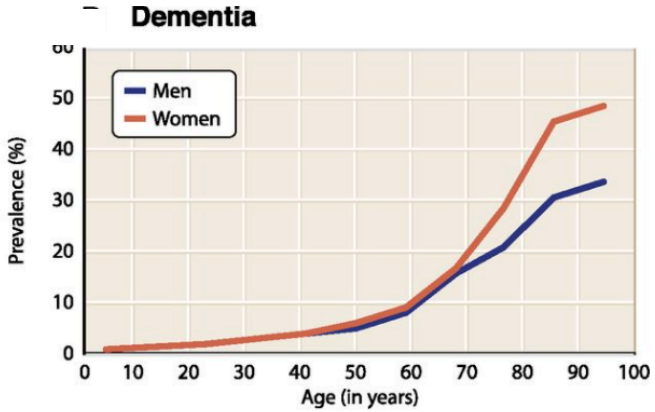
Outline

- **Recap**: intro to the problem, EDA, and preprocessing
- **Cross-validation** (CV): pipeline, models and parameter tuning
- **Results**: select, evaluate, and interpret the best model
- **Outlook**: to improve the model

Recap: aging is one of the biggest disease risk factors, but it is hard to evaluate anti-aging interventions



UN World population Prospect (2008)



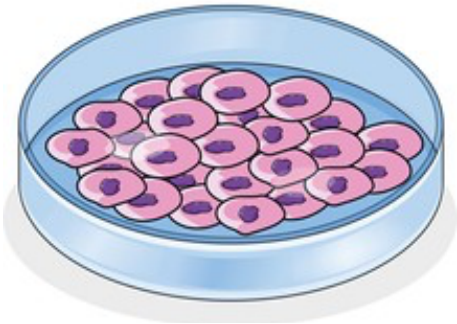
Luu, Jennings, and Krzysztof Palczewski. *PNAS* (2018)

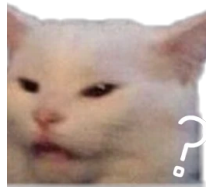
Maximum Lifespan

4 years

40 years

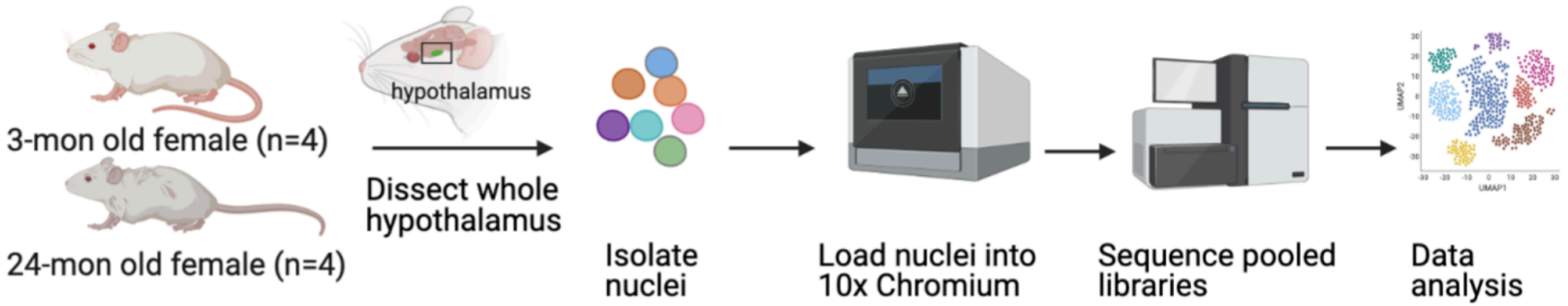
122.5 years





Maegawa et al. *Nat communications* (2017)

Recap: classification based on single cell RNAseq data



Hajdarovic, Kaitlyn H., et al. "Single cell analysis of the aging hypothalamus." under revision

25,002 nuclei/cells **Gene names**

features	Xkr4	Gm1992
AAACCTGAGACTAGAT-1_1	1.540812	0.000000
AAACCTGAGGCCCGTT-1_1	0.000000	0.000000

Cell ID

Additional features related to total gene expression

sum	x_sum	x_prop
3477.047641	128.435941	0.036938
2213.750159	74.358735	0.033589

tree.ident	**Target**
Avp/Oxt	Young (0)
Nrg1/Nnat	Aged (1)

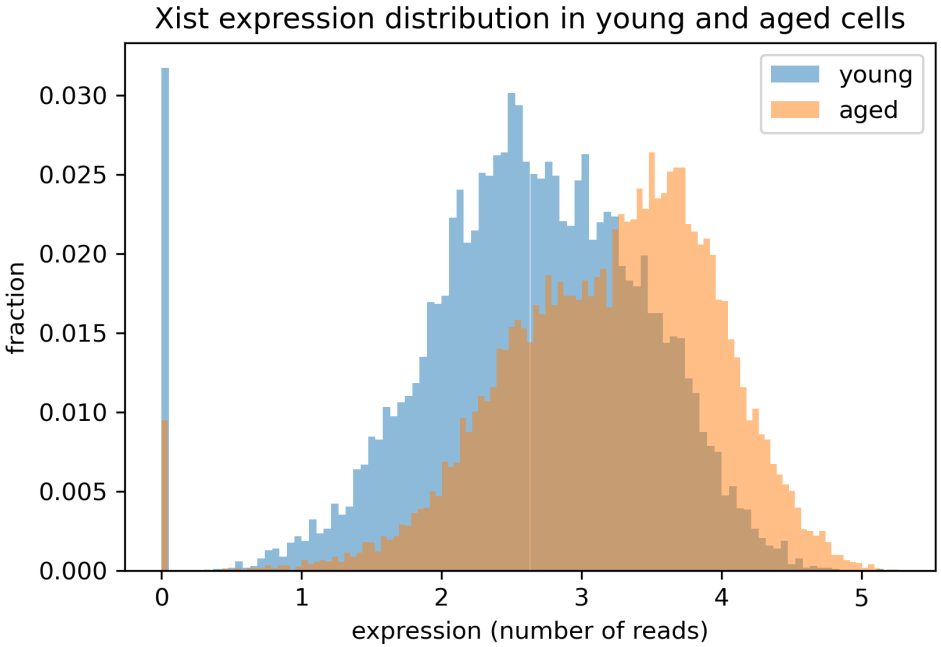
281 Numerical features

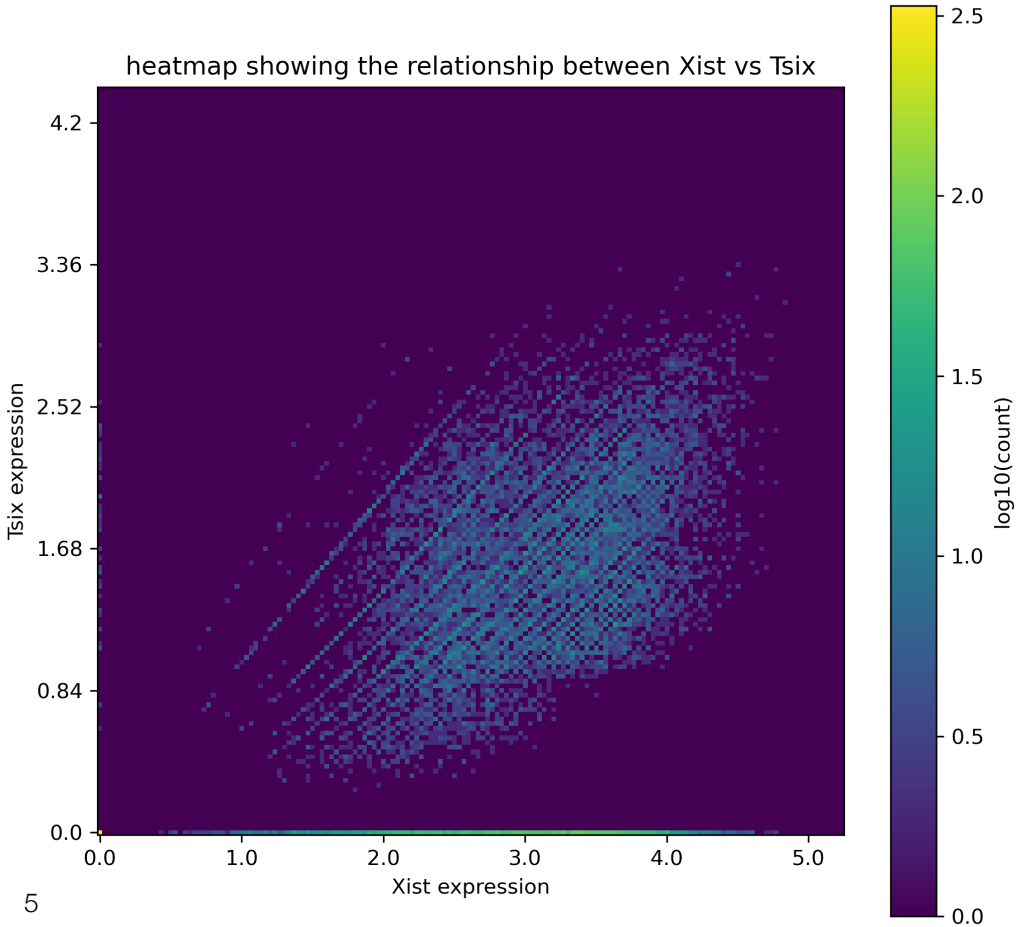
Not bounded, zero-inflated negative binomial—> **StandardScaler**

1 Categorical feature with 34 categories

not ordered —> **OneHotEncoder**

Recap: increased expression of X-genes *Xist* and *Tsix* during aging





Cross validation: KFold splitting (GridSearchCV)

Splitting:
- To predict cells within the previously investigated animals: IID, not time-series
- train-validation-test: 64-16-20
- train_test_split (20% for testing) + KFold (5 folds) to ensure no overlap between two folds

Preprocessing:
- 281 numerical features (unbounded, continuous): StandardScaler
- 1 categorical feature (unordered): OneHotEncoder

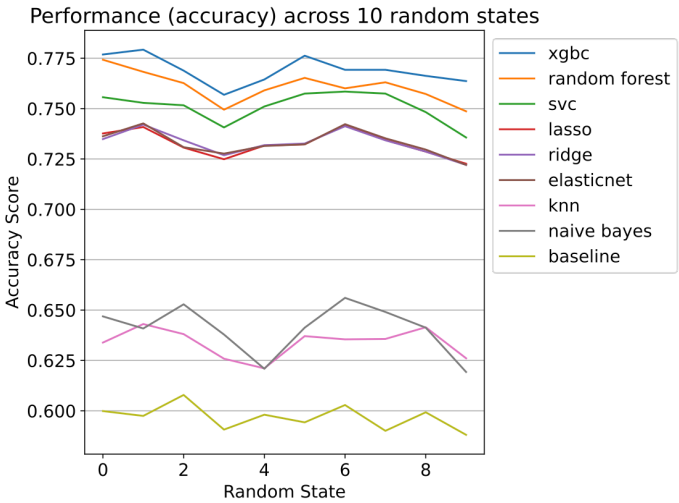
GridSearchCV for each model:
- pipeline with KFold CV to avoid data leakage
- accuracy score (not imbalanced)

Loop through**10 random states**:
- measure uncertainties due to splitting and non-determistic models
- return best hyperparameters, validation and test scores for each state

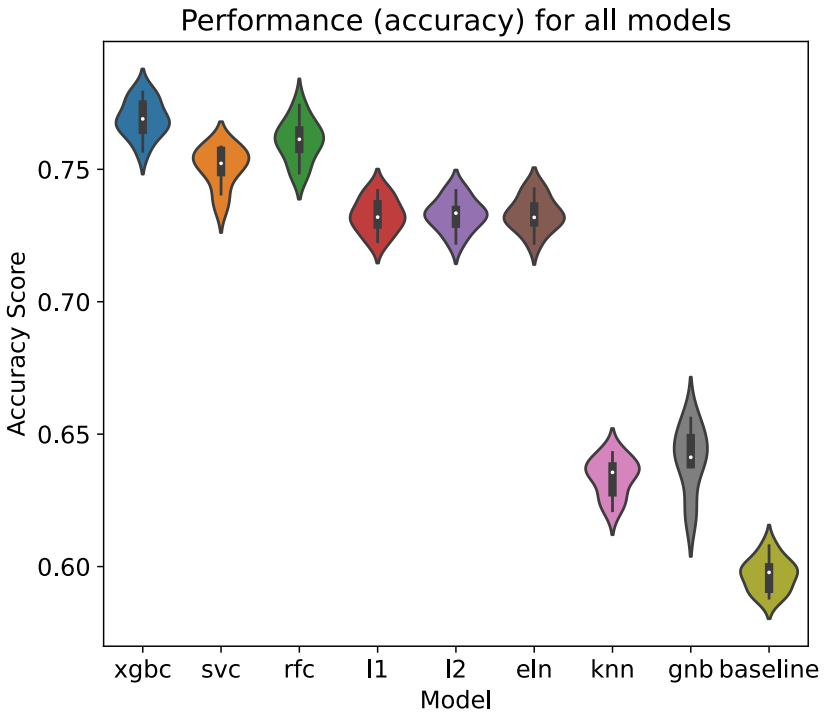
Cross validation: models and parameters tuned

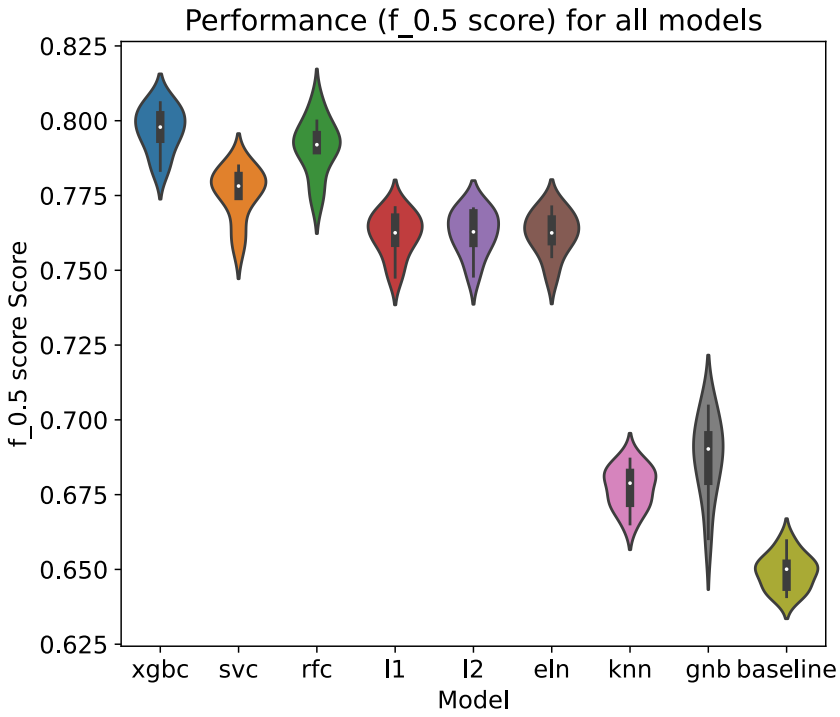
Tabel 1. Parameters used for tuning models

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



Results: model performance (test scores) — XGBoost Classifier outperforms others

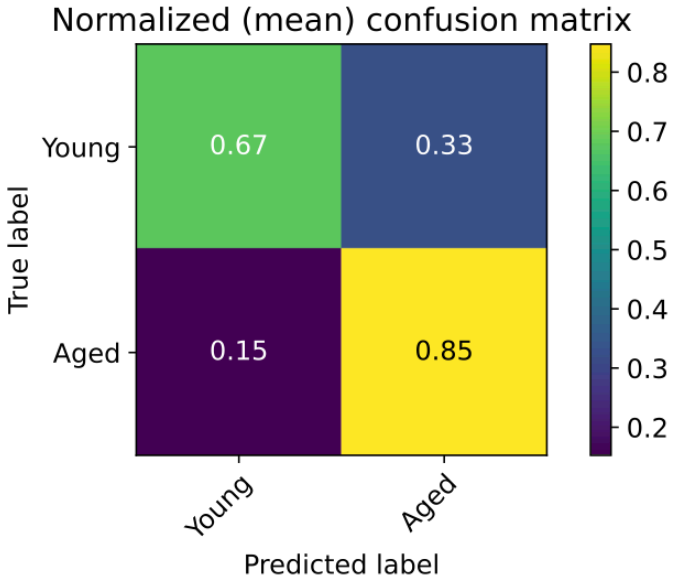




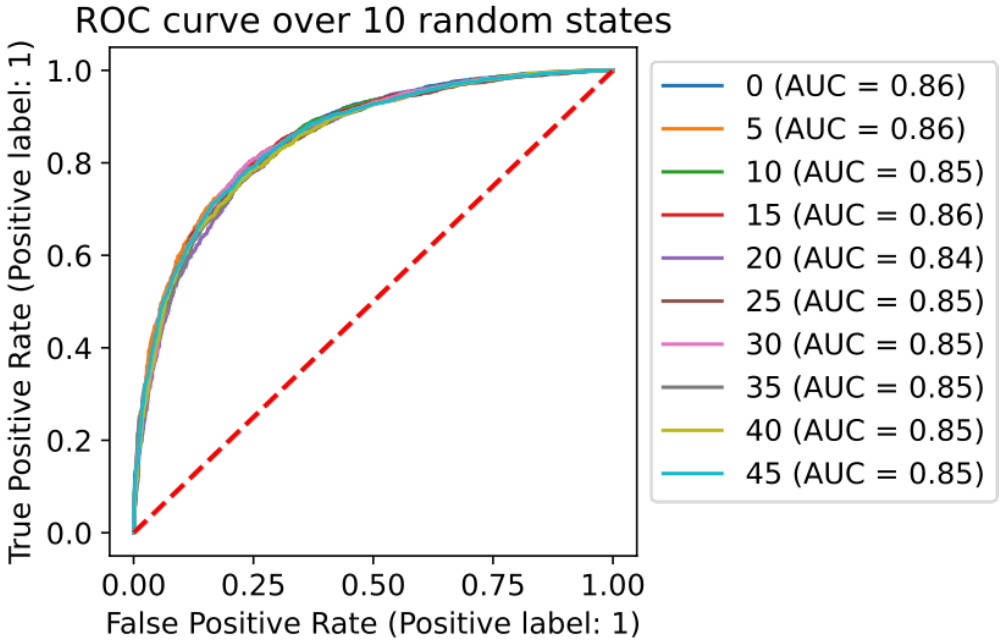
f scores with beta=0.5 (put more emphasis on precision since it is expensive to perform the anti-aging interventions)

Results: retrained the XGBoost with the selected parameters over 50 random states

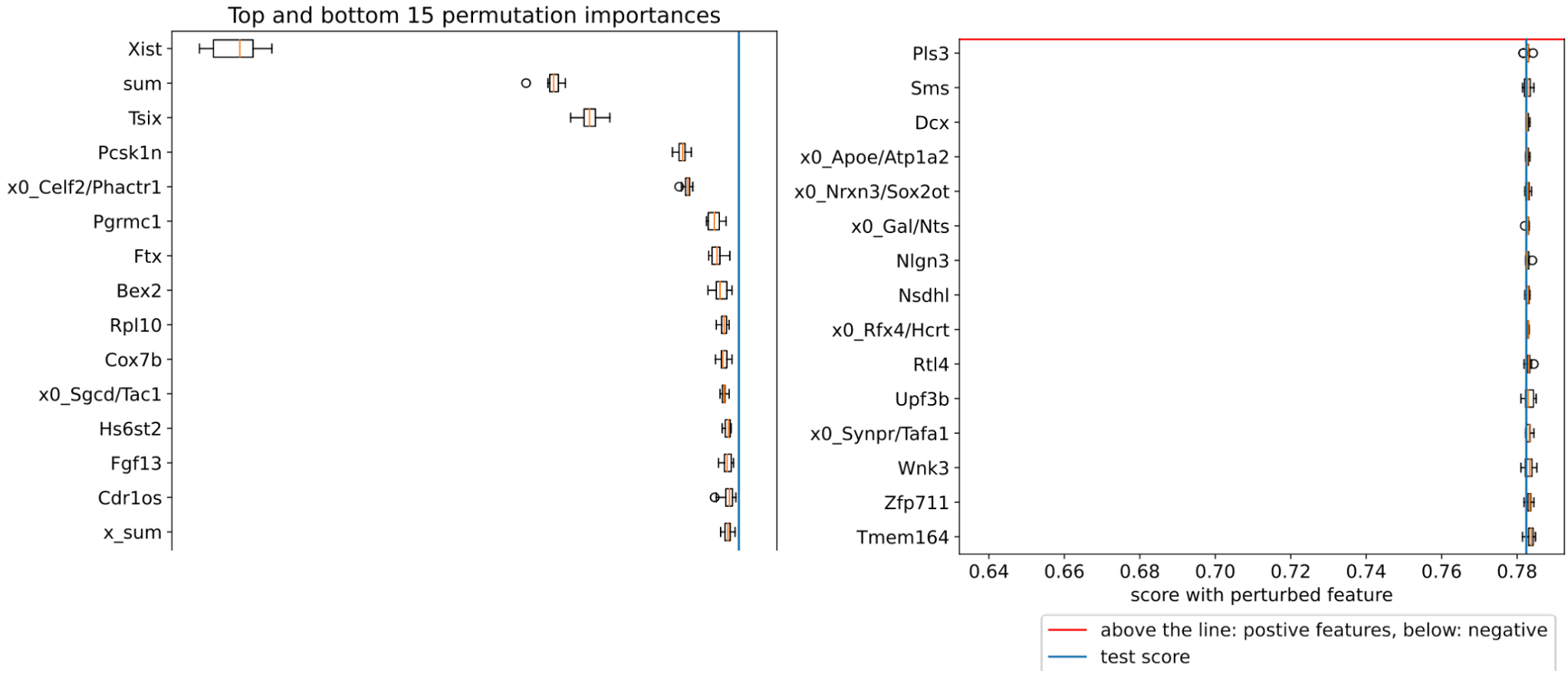
- max_depth = 5 with early_stop
- baseline accuracy: 0.596 ± 0.007
- model accuracy: 0.778 ± 0.006



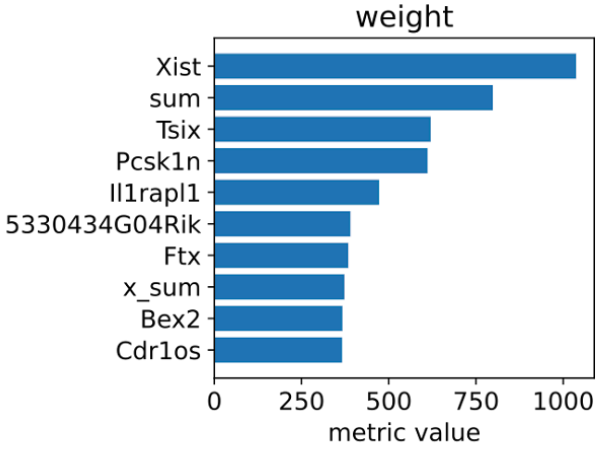
consistent over time

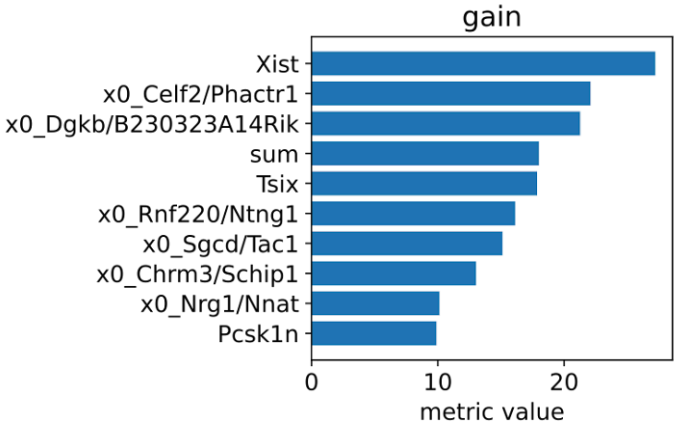


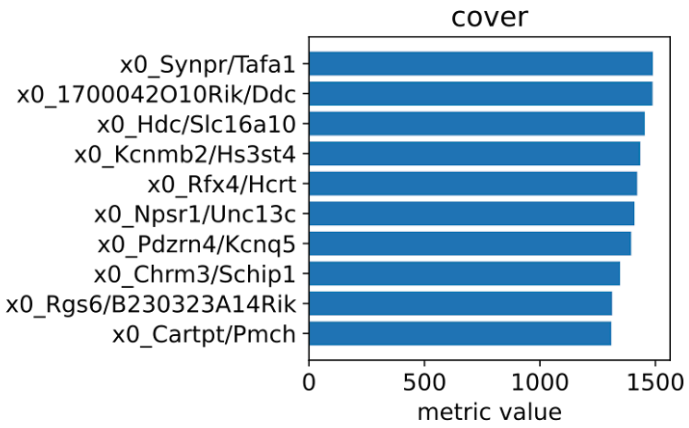
Results: permutation feature importance

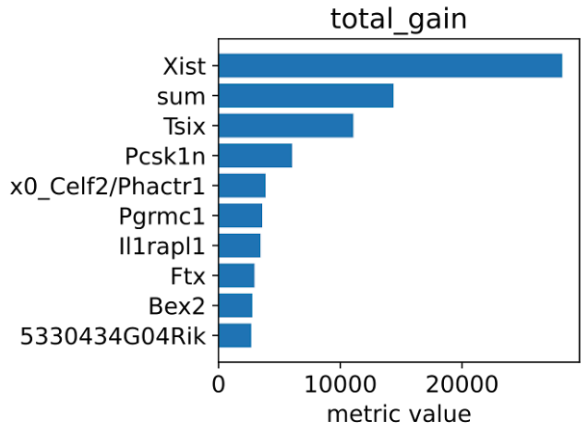


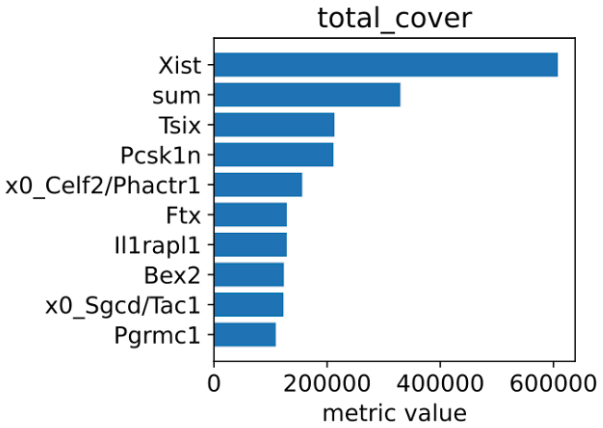
Results: important features in XGBoost metrics



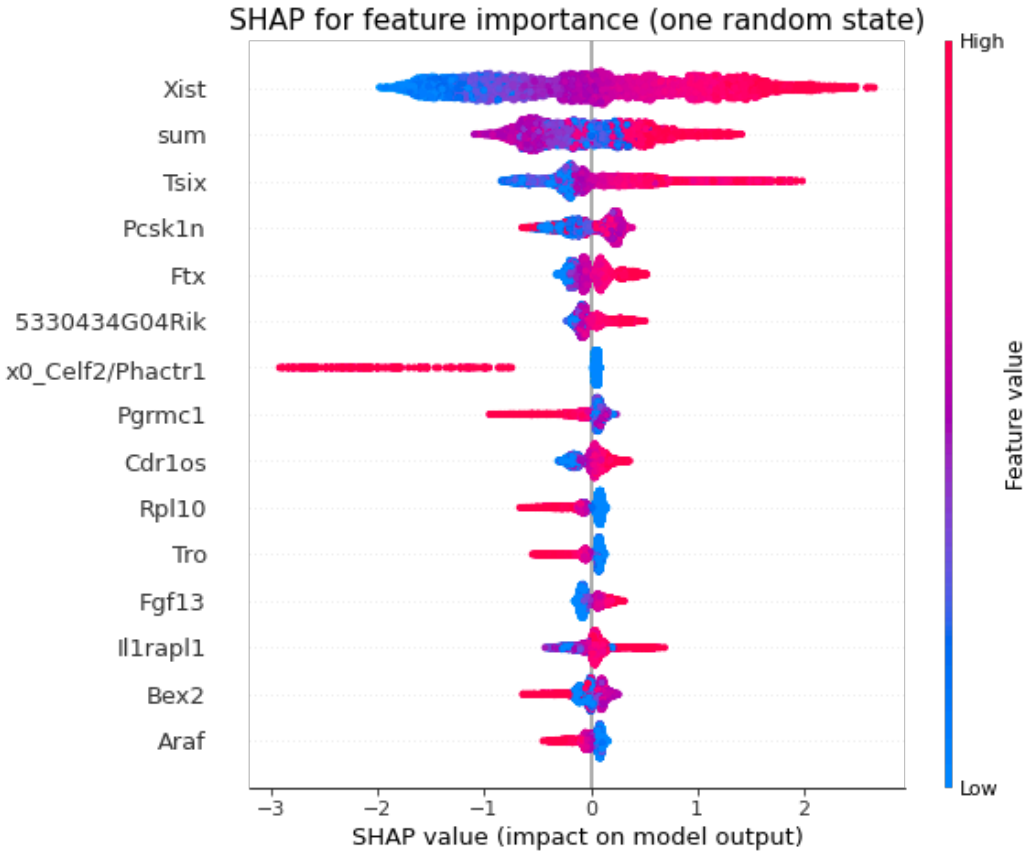




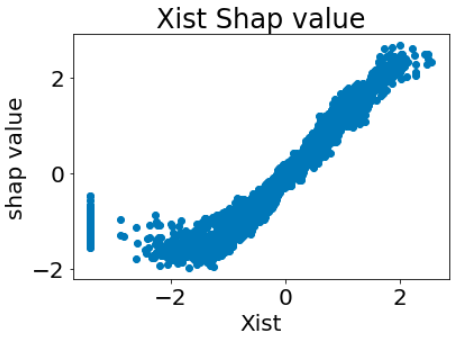


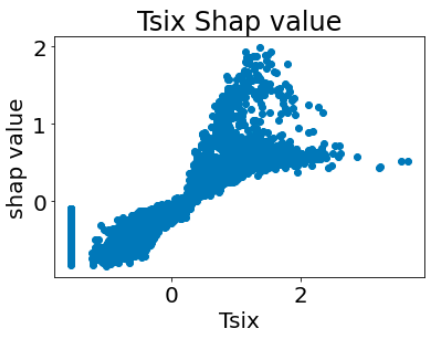


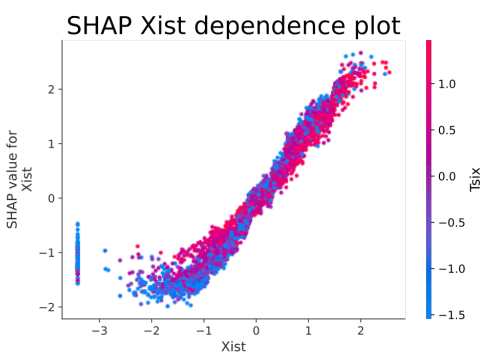
Results: SHAP feature importance for global and local interpretability

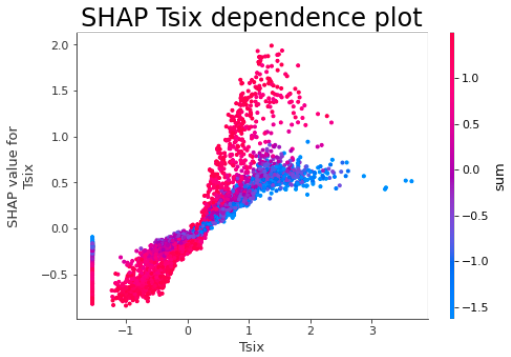


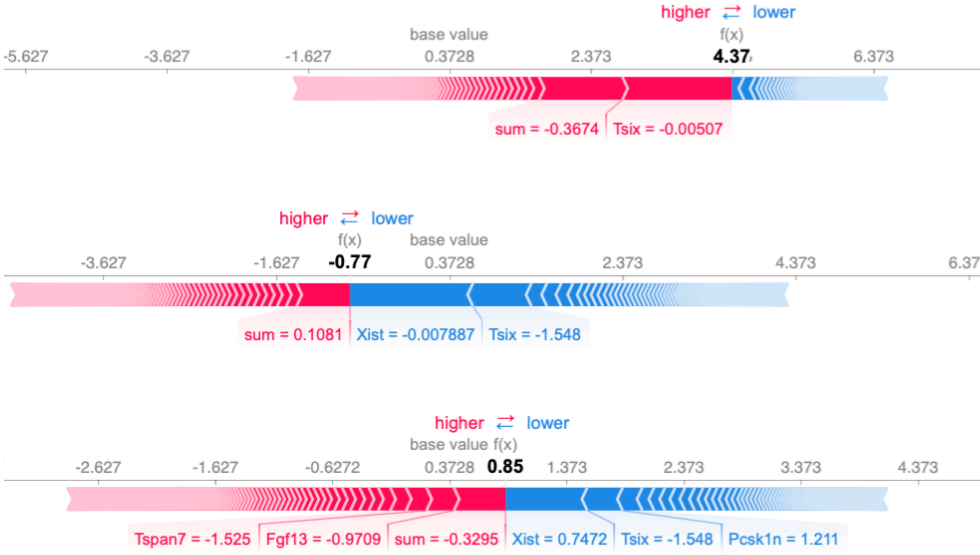
Results: SHAP local features



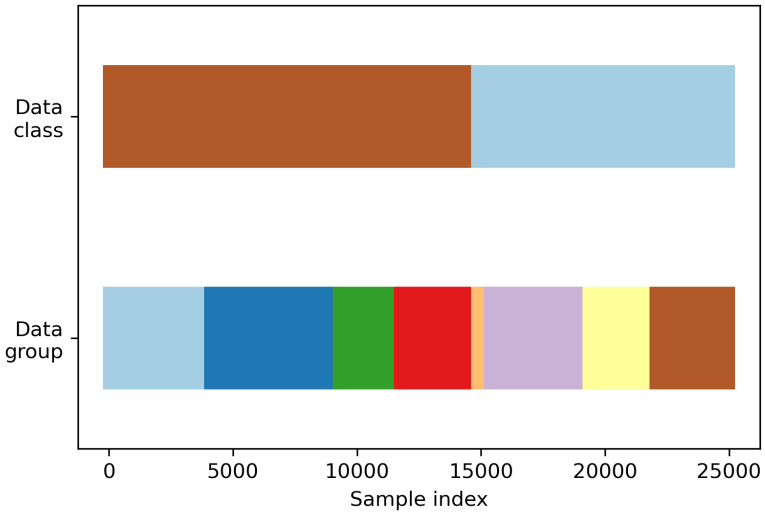








Outlook



- GroupKFold splitting for broader applications, together with more animals or ages
- XGBoost linear models to improve interpretability
- For better prediction: booster parameters like gamma could be further tuned, together with gpu implementation.
- Adding more features (not just X chromosome genes)
- With more animals in different time points, probably change the classification into a regression (specific age)