Session 3: Machine Learning Analysis

Resources

Discord

QA

Support

Interact

Amazon SageMaker

Link

Raw data

Analysis code

Visualization

GitHub

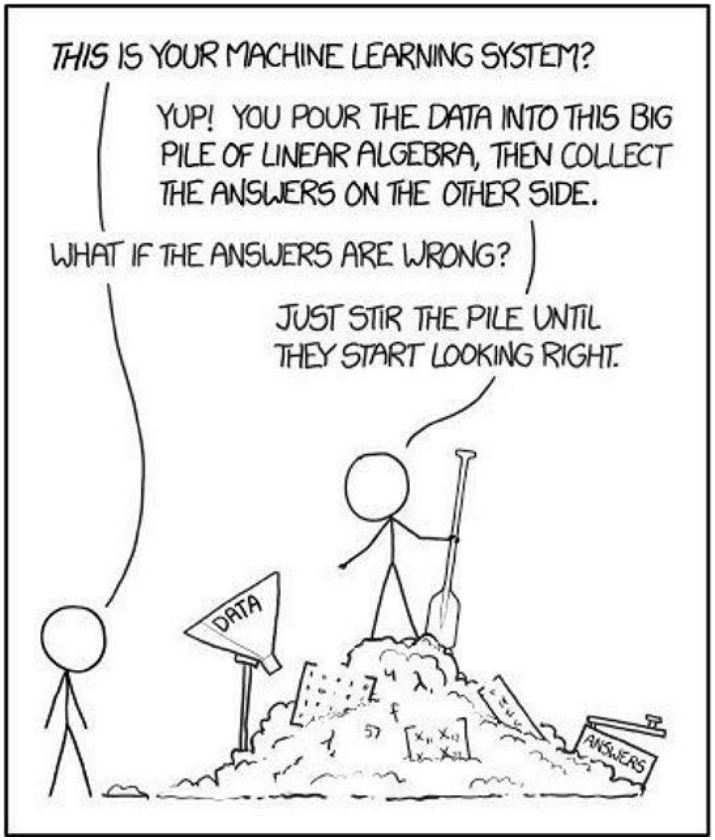
Resources

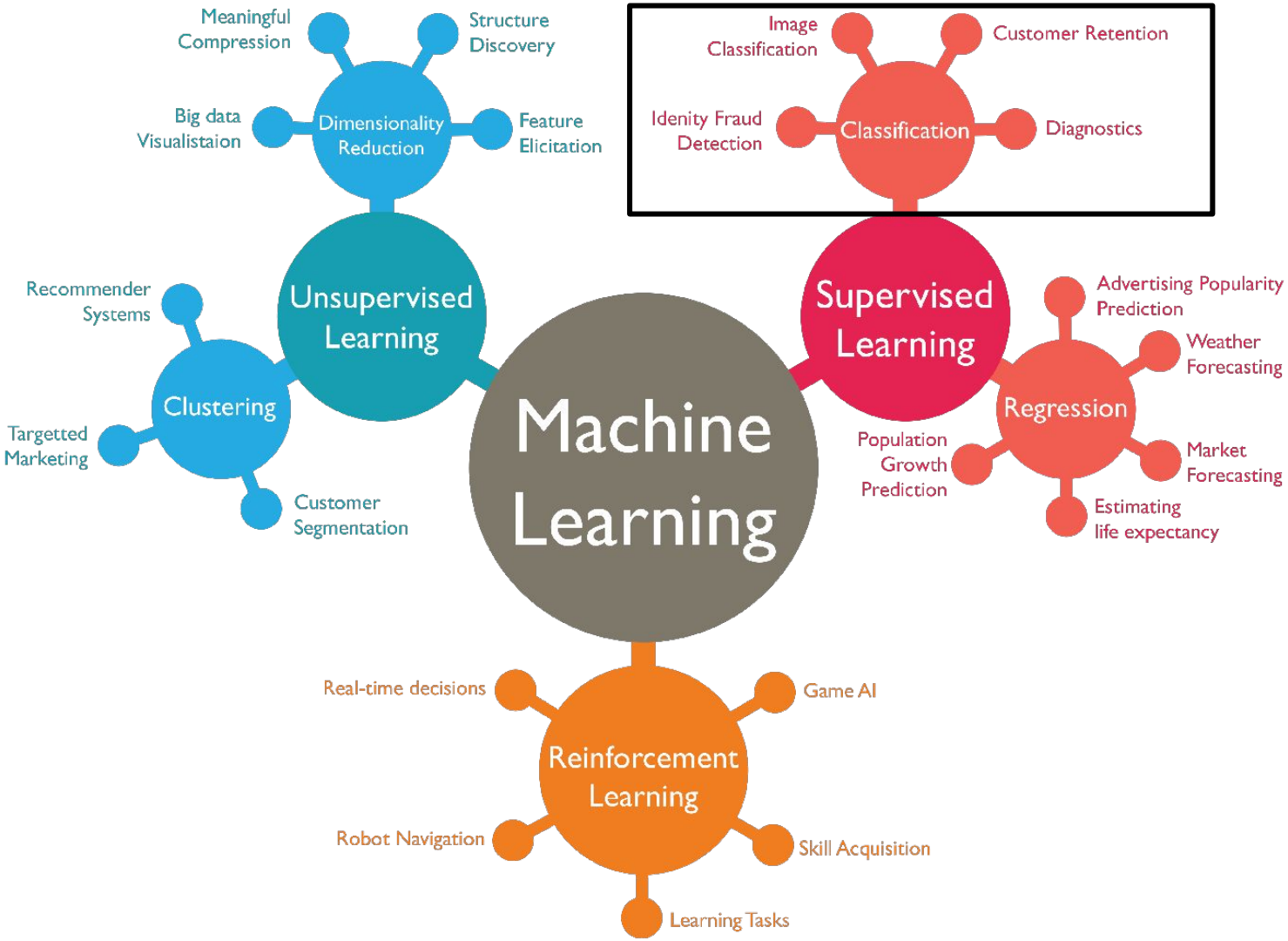
Slides

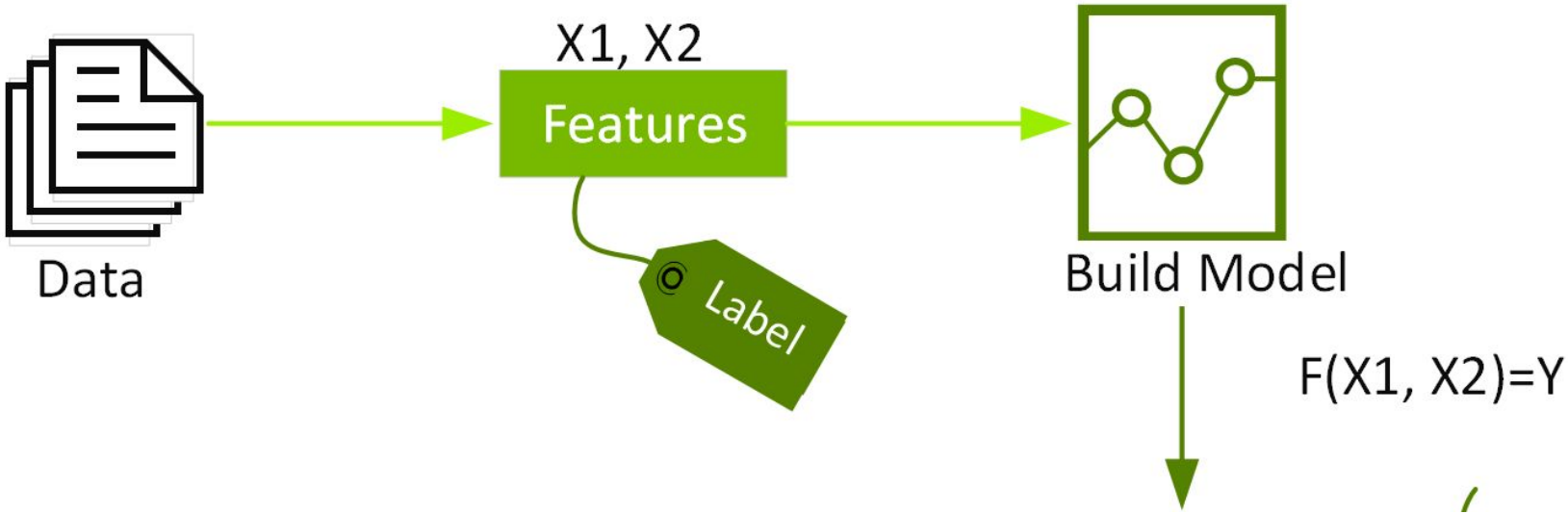
Notebooks

https://github.com/PriceLab/Aging_Workshop_24

“Remember, the other team is using Machine Learning on your games to predict your play. So, kick the ball with your other foot!”




Machine Learning
Unsupervised Learning
Meaningful Compression
Structure Discovery
Dimensionality Reduction
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Image Classification
Customer Retention
Classification
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating life expectancy
Population Growth Prediction
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks


X1, X2
Features
Data
Label
Build Model
F(X1, X2)=Y

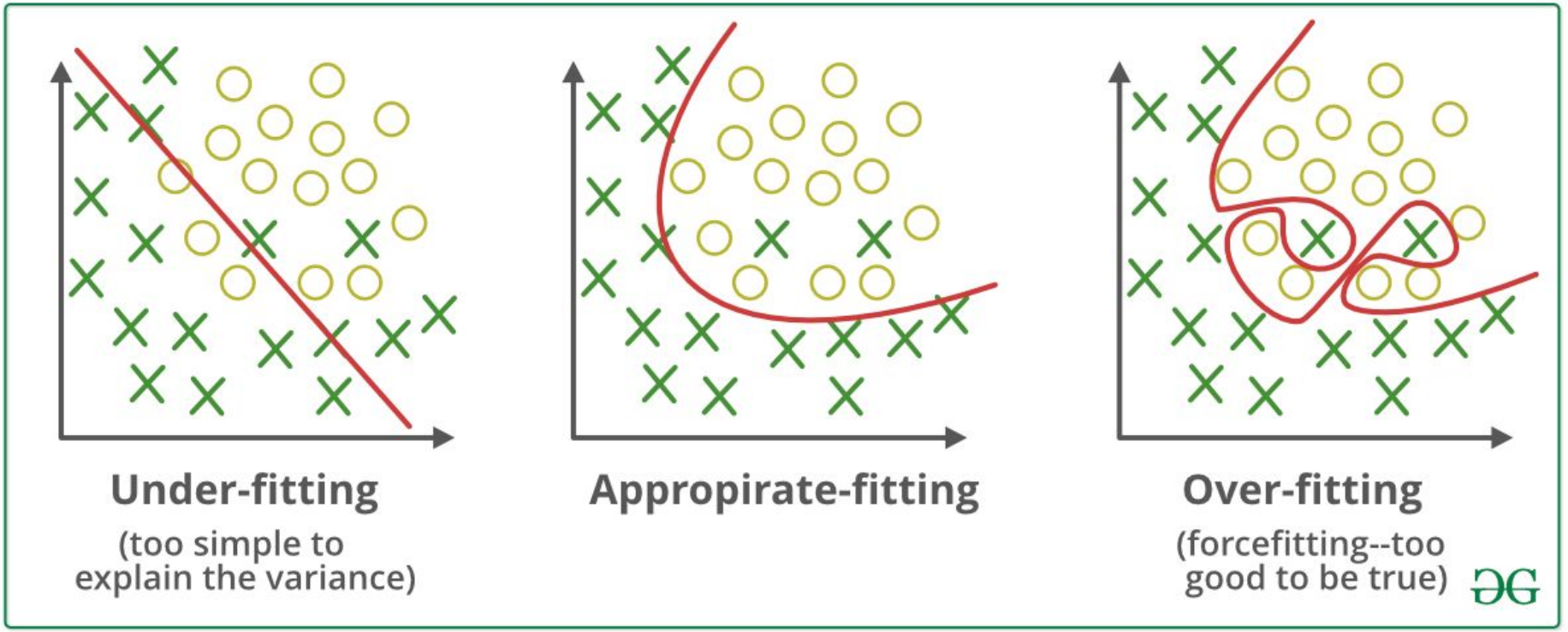
“Essentially, all models are wrong, but some are useful”

-George Box

Steps

1) Define the problem & Choose your outcome
2) Choose a dataset
3) Decide which variables to include
4) Pick a model
5) Train & test your model(s) and measure performance

Model Fitting



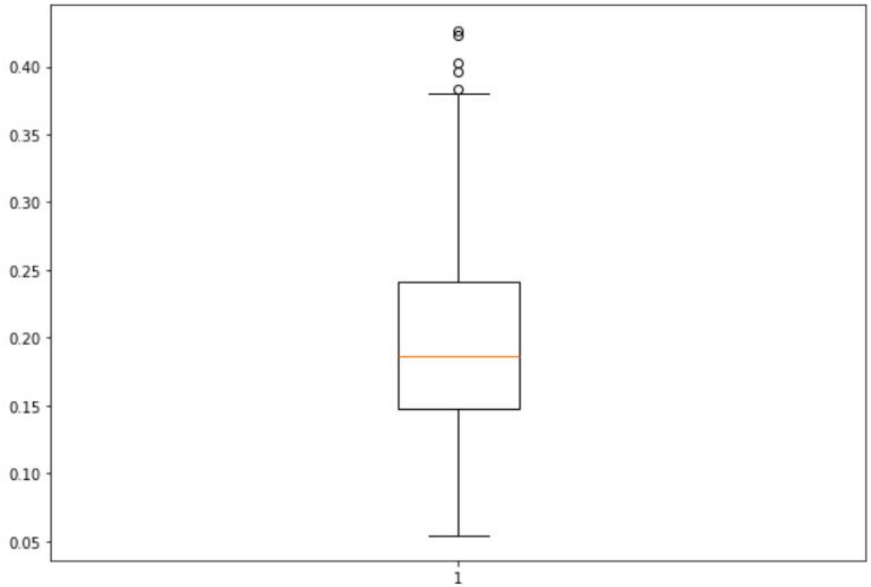
Source

Step 1: Defining our Problem

Step 1a: What is the problem we're trying to solve?

Step 1b: What is the outcome we're trying to predict?

- Outcome Definition: High vs. low Frailty Index

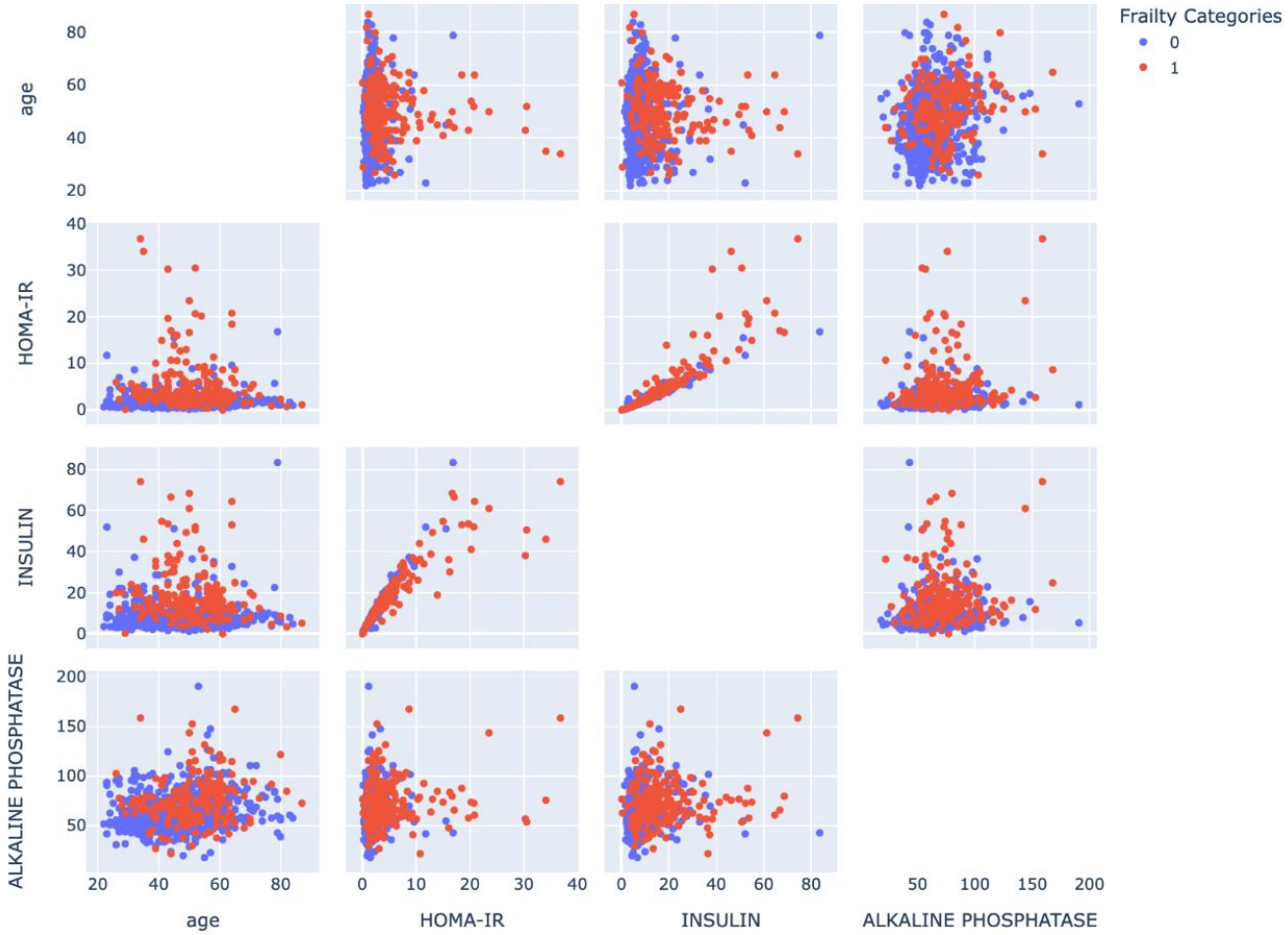


Step 2: Picking our Dataset

Patient	Age	Sex	Glucose	Triglycerides	Outcome
001	25	F	75	80	0
002	67	F	100	175	1
003	44	M	50	66	1

- 769 patients
- 76 features
  - Demographics (age, sex)
  - WGCNA (metabolomics and proteomics)
  - Labs
- Outcome: Frailty Index (upper quartile)

Data Analysis/Exploration



Step 3: Which features to include?

- Domain-informed
- Subset selection
  - Best subset: All possible models → computationally infeasible when p>40
  - Forward stepwise selection: Only add the feature that gives us the best “next” model
  - Backwards stepwise selection: Start with a model containing all predictors and remove them one at a time by seeing which feature has the least effect on the model performance
  - Choose best model out of all the models → on the **test set!**
- **Shrinkage**
  - **Lasso and Ridge**
    - *Shrinkage penalty* (λ) that penalize larger coefficients
- Dimension reduction
  - PCA

Feature Engineering

- Add additional features
  - BMI
  - Education
  - Income
  - Medications
- Remove highly correlated variables
- Missingness
- Choose a different outcome: Predict age
- Class imbalance

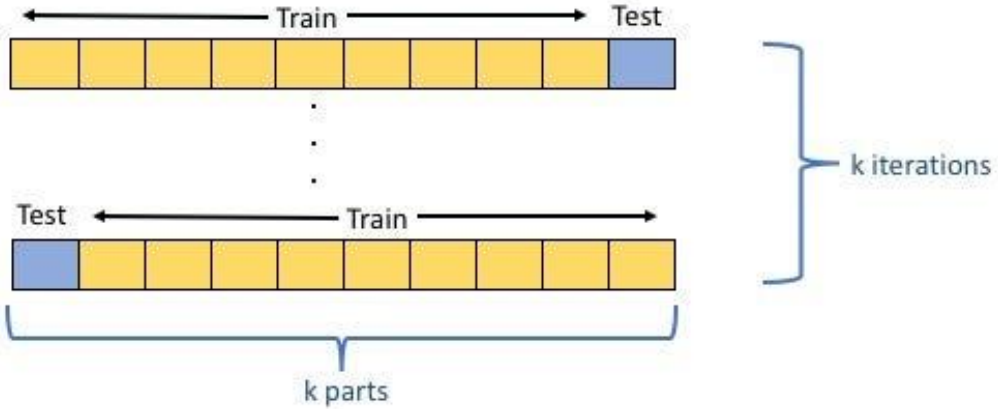
Step 4: Choosing a model

- **Logistic Regression**
- Tree-based models
  - Random Forest
  - **XGBoost**
- Support Vector Machines
- Neural Networks

Step 5a: Train & test your model(s)

- First: Split your data into training (80%) and testing (20%)
- Next (optional but recommended): Do cross-validation on the **training data**

1. Divide the sample data into k parts.
2. Use k-1 of the parts for training, and 1 for testing.
3. Repeat the procedure k times, rotating the test set.
4. Determine an expected performance metric (mean square error, misclassification error rate, confidence interval, or other appropriate metric) based on the results across the iterations

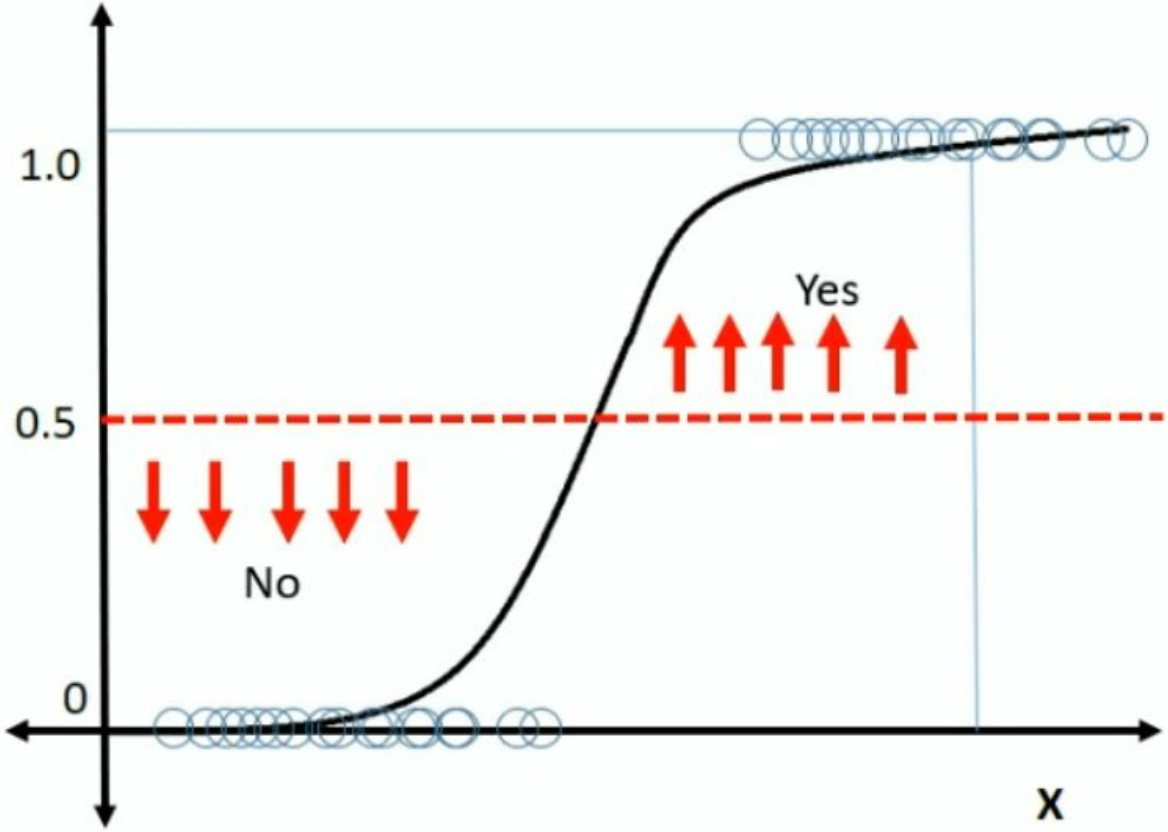


Step 5b: Measure performance

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

3	1
1	3

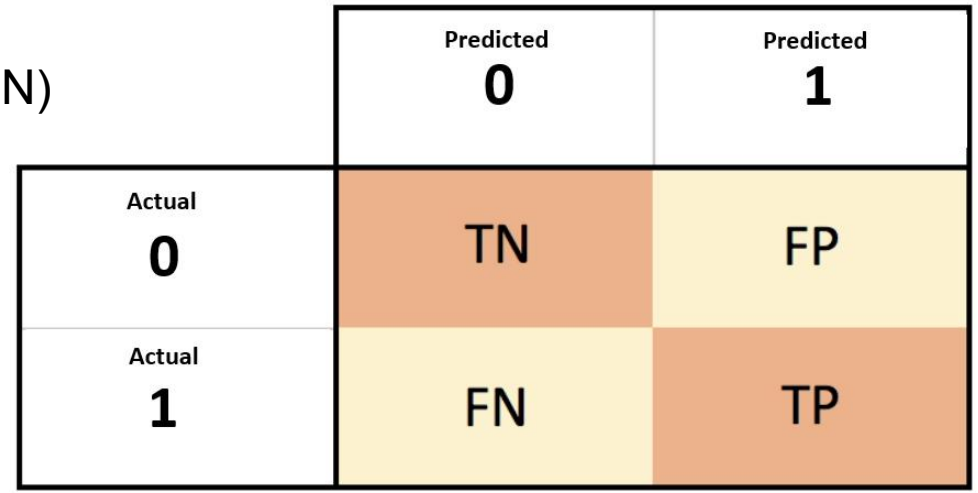
Probability	Predicted	Actual	Label
0.25	1	0	TN
0.4	0	1	FN
0.89	1	1	TP
0.65	1	1	TP
0.2	0	0	TN
0.3	1	0	TN
0.55	1	0	FP
0.8	1	1	TP



Threshold Value of 50% probability
> 0.5 Success
< 0.5 Failure

Performance Metrics

- Precision (TP/TP + FP)
- Recall (Sensitivity) (TP/TP + FN)
- Specificity (TN/TN+ FP)
- F1
- AUROC
- PR AUC

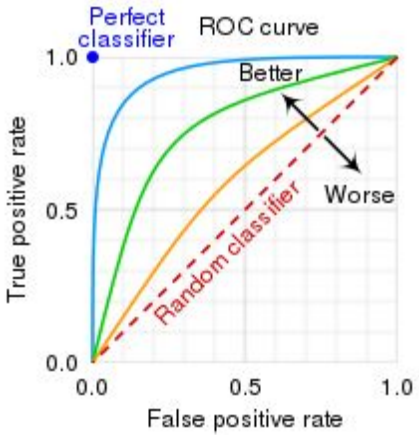


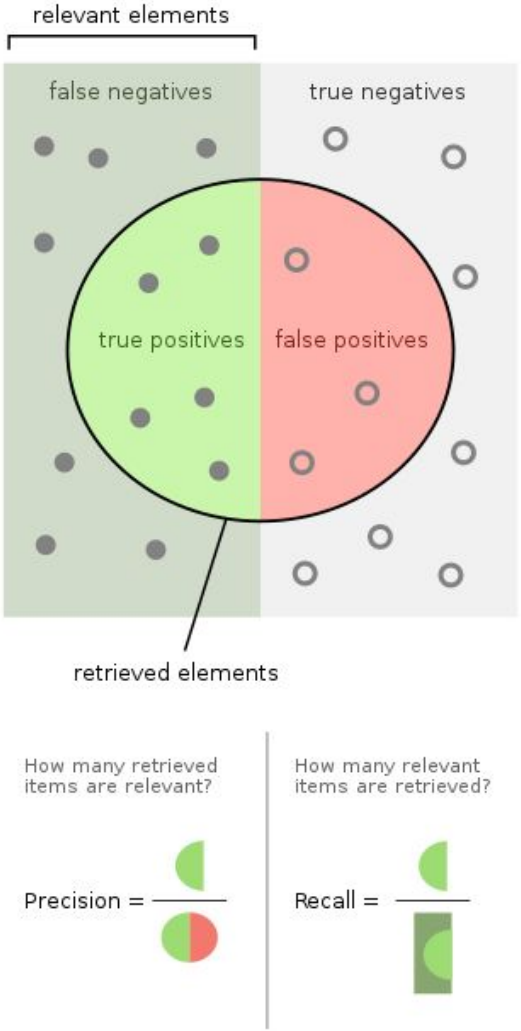
Precision: 9/20 = 0.45
Recall: 9/10 = 0.9
Specificity: 12/23 = 0.52

12	11
1	9

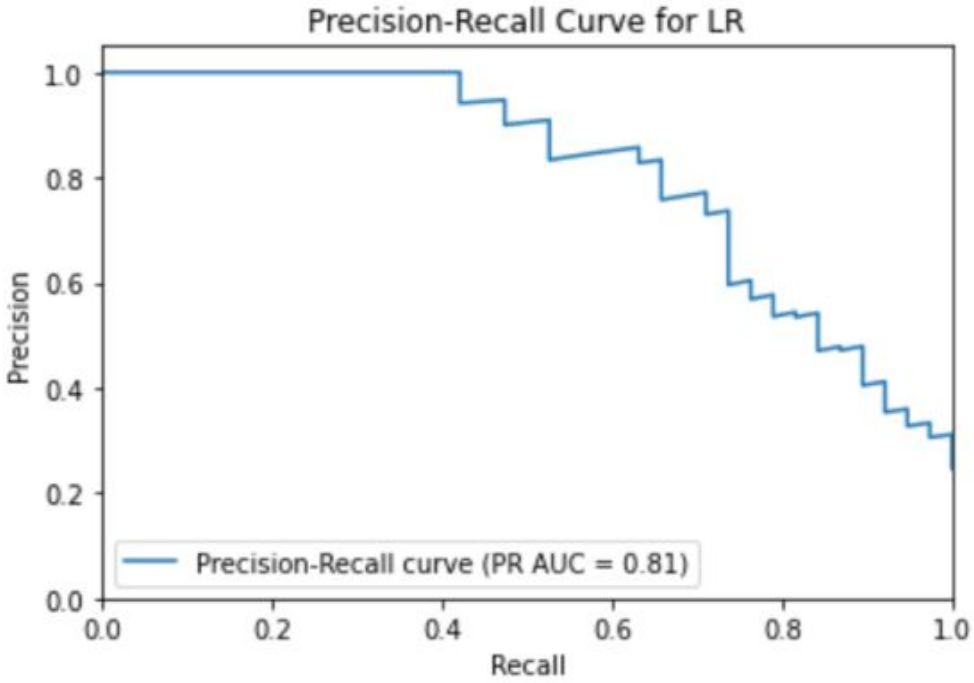
Performance Metrics

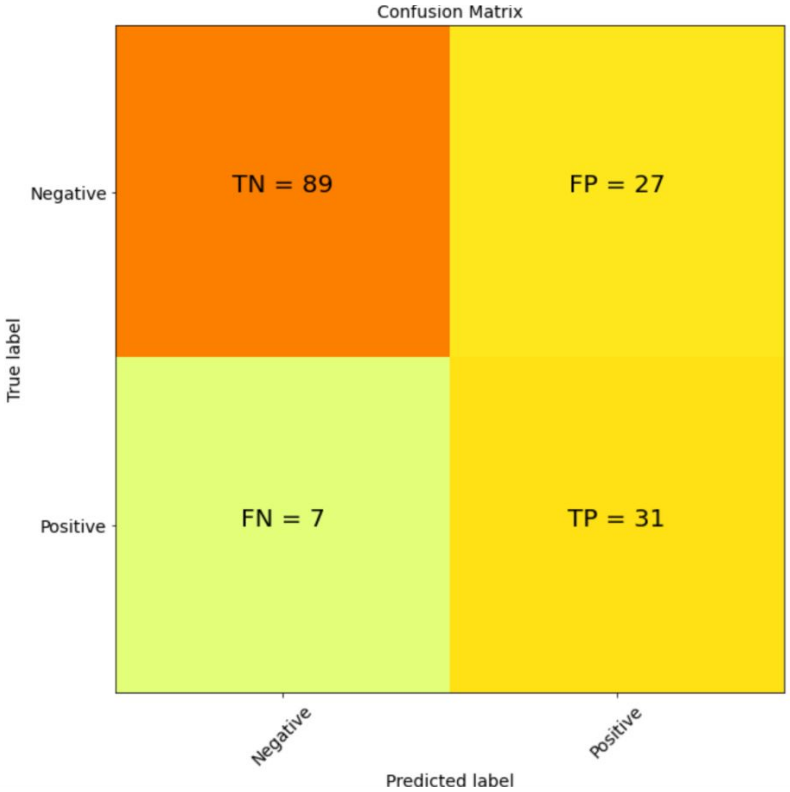
- Precision (TP/TP + FP)
- Recall (Sensitivity) (TPR) (TP/TP + FN)
- Specificity (TN/TN+ FP)
- F1
- ROC AUC (AUROC)
- PR AUC





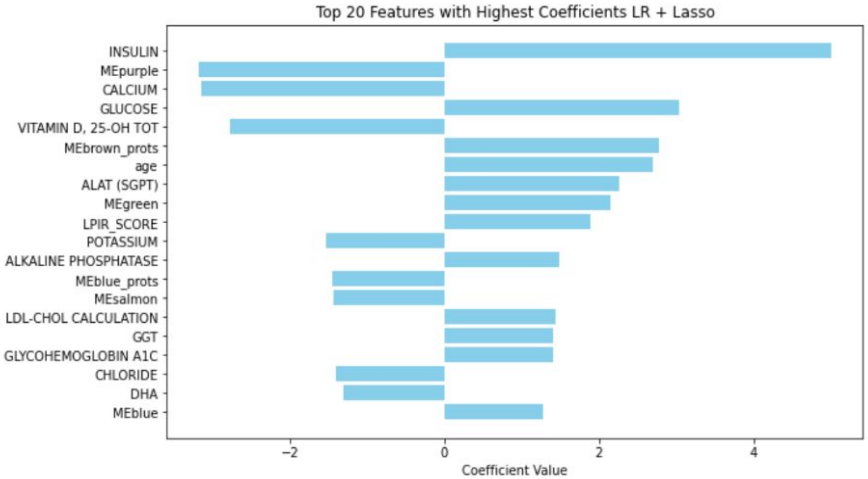
Results: Logistic Regression on the Test Set (154 Patients)





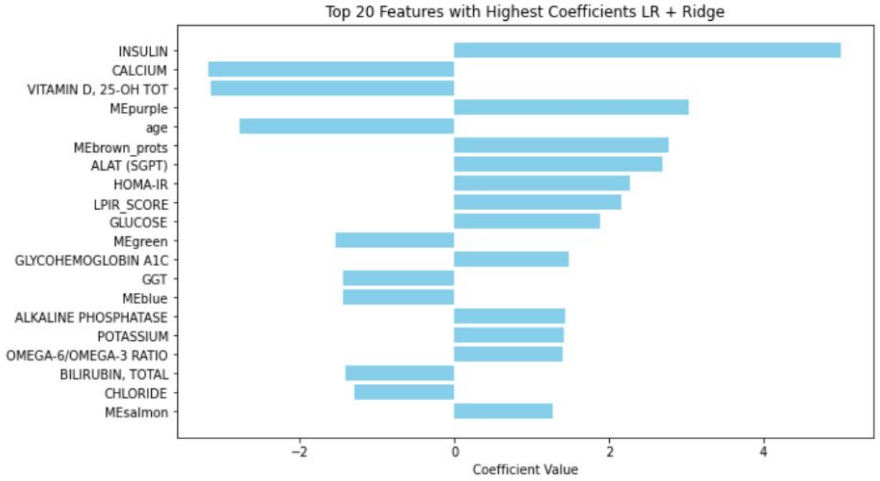
Lasso

PR AUC: 0.8025563981672981
ROC AUC: 0.8788566243194192
Accuracy LR + Lasso: 0.7792207792207793
Precision LR + Lasso: 0.5344827586206896
Recall LR + Lasso: 0.8157894736842105
F1 Score LR + Lasso: 0.6458333333333334

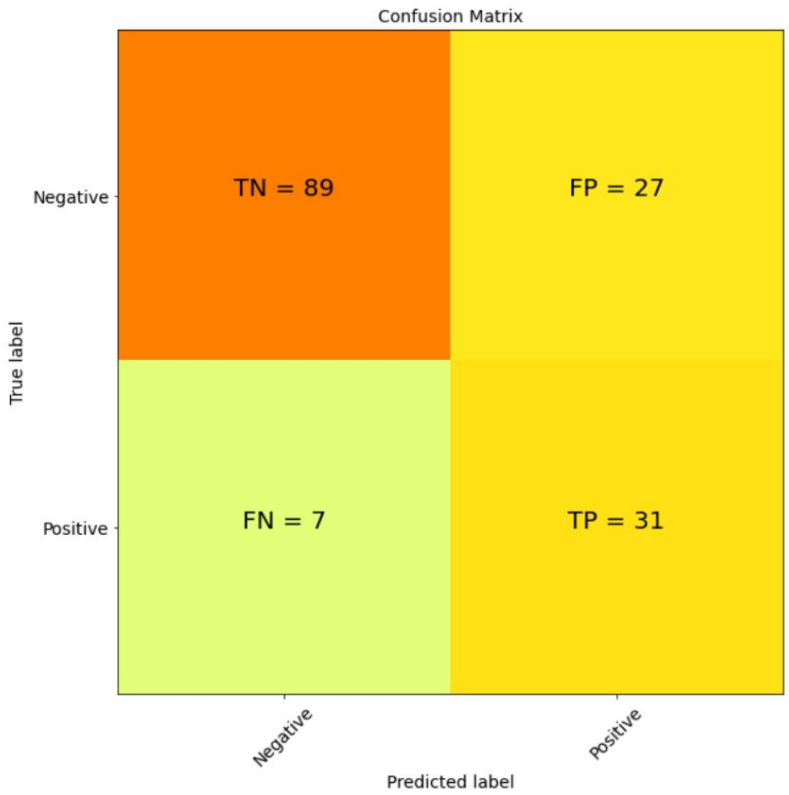


Ridge

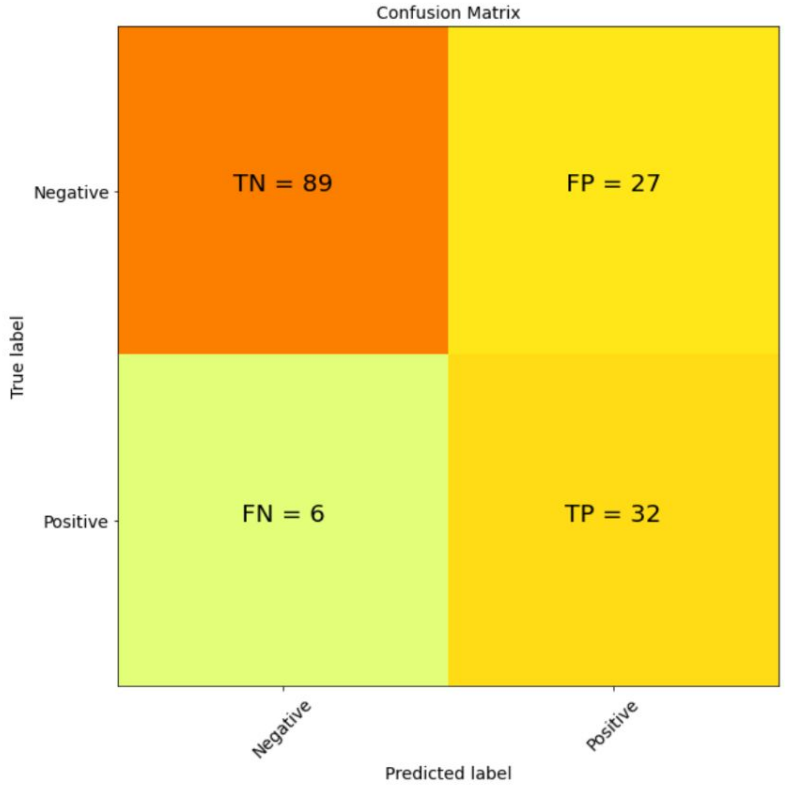
PR AUC Ridge: 0.8090447827152129
ROC AUC Ridge: 0.889065335753176
Accuracy LR + Ridge: 0.7857142857142857
Precision LR + Ridge: 0.5423728813559322
Recall LR + Ridge: 0.8421052631578947
F1 Score LR + Ridge: 0.6597938144329897



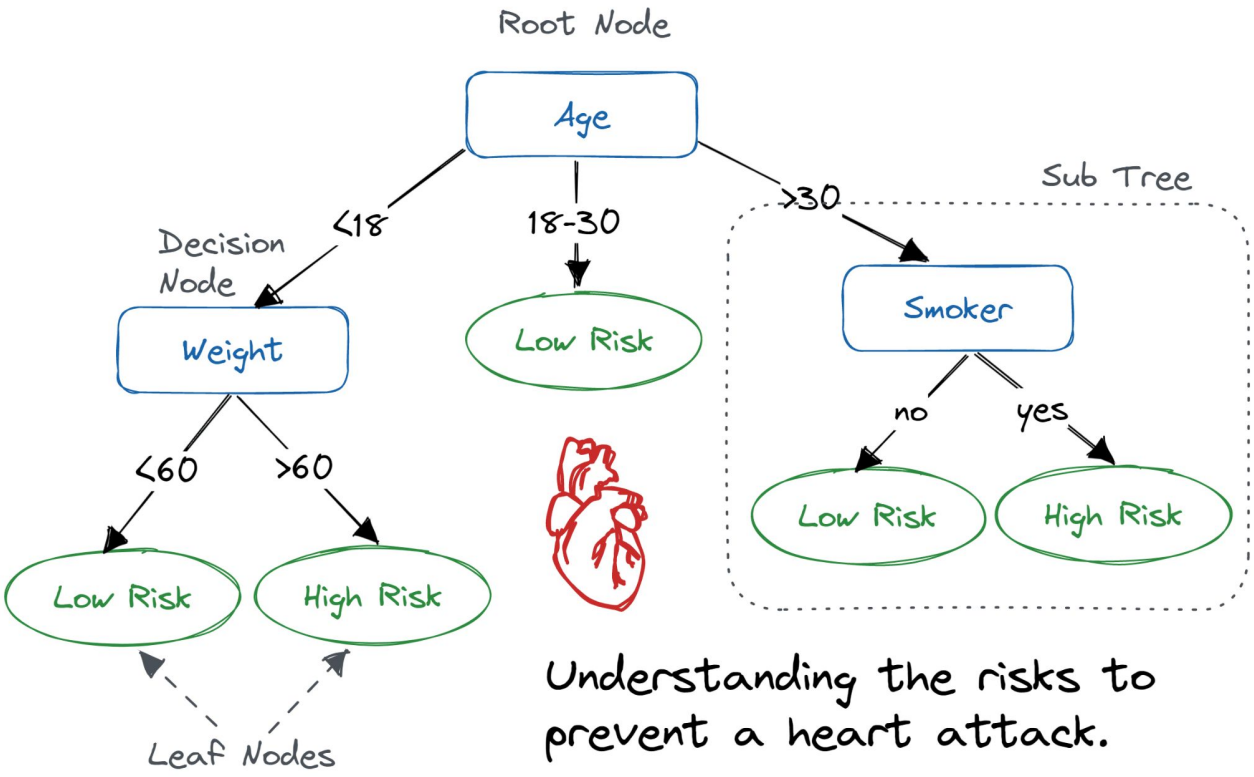
Lasso


Confusion Matrix
TN = 89
FP = 27
FN = 7
TP = 31
Negative
Positive
True label
Negative
Positive
Predicted label

Ridge


Confusion Matrix
TN = 89
FP = 27
FN = 6
TP = 32
Negative
Positive
True label
Negative
Positive
Predicted label

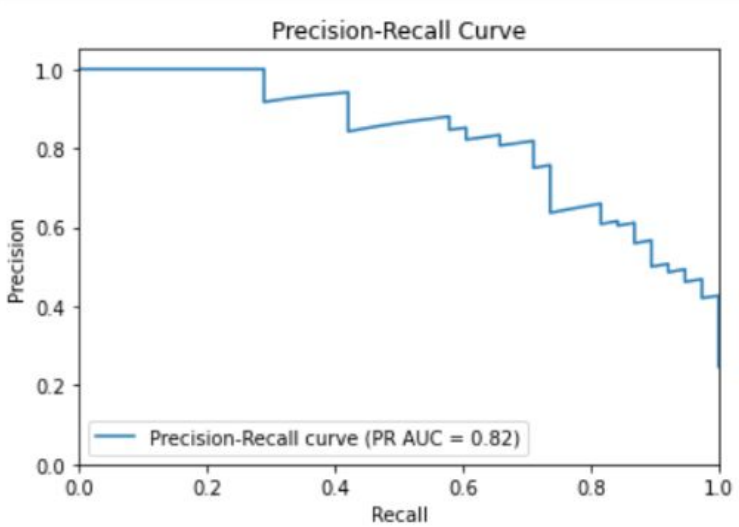
Tree-based methods

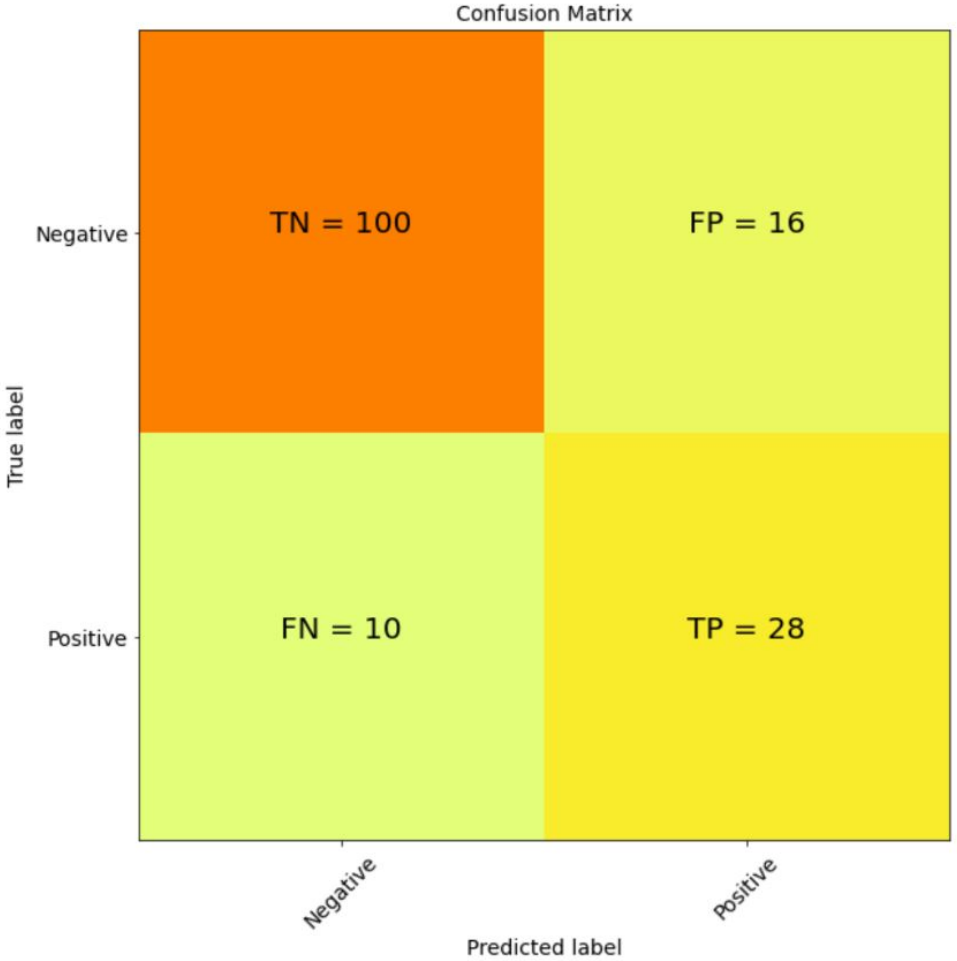


Source

XGBoost

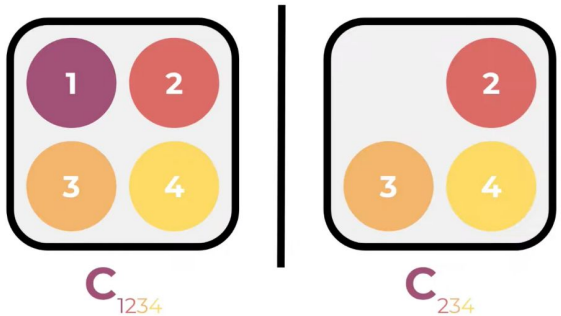
PR AUC: 0.8241714305543074
ROC AUC: 0.9235480943738656
Accuracy XGB: 0.8311688311688312
Precision XGB: 0.6363636363636364
Recall XGB: 0.7368421052631579
F1 Score XGB: 0.6829268292682927



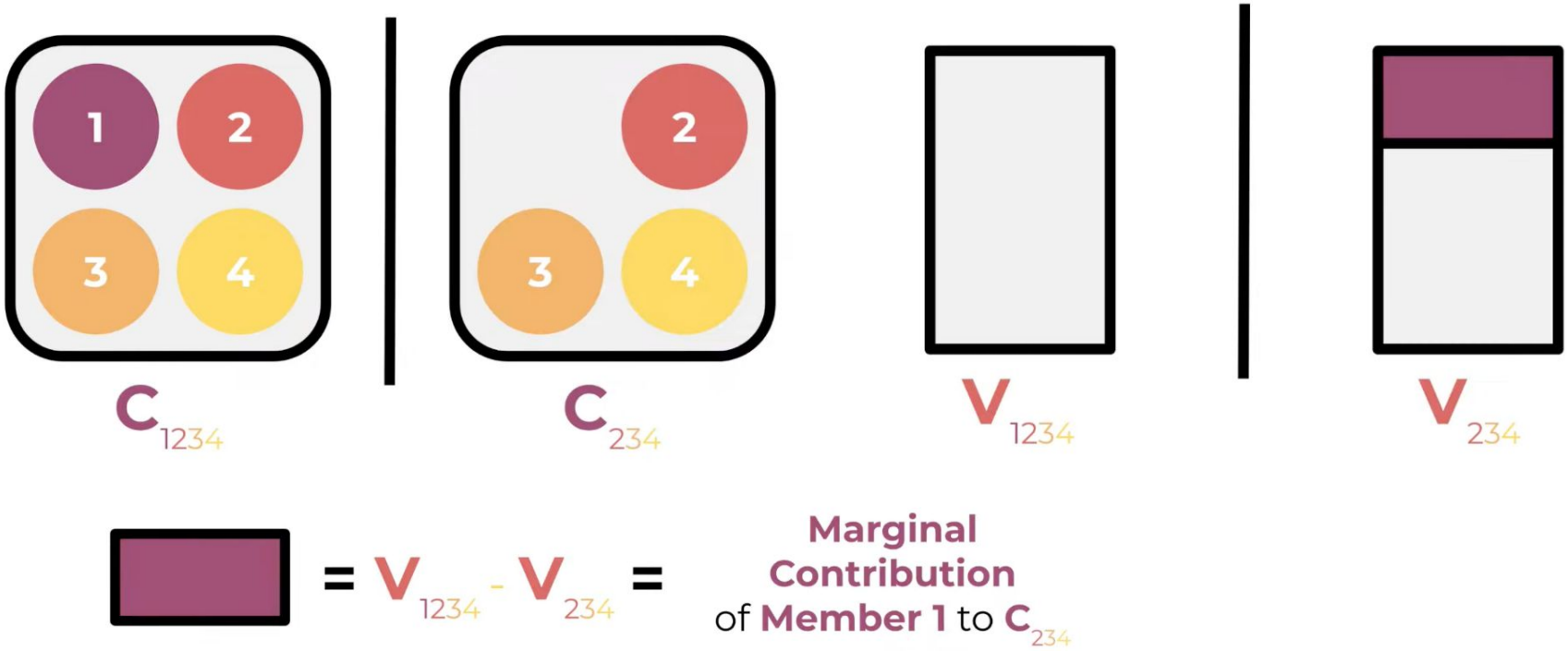

Confusion Matrix
TN = 100
FP = 16
FN = 10
TP = 28
Negative
Positive
True label
Negative
Positive
Predicted label

SHAP (SHapley Additive exPlanations)

- Based on Shapley Values from Game Theory
- If we have a coalition **C** that collaborates to produce a value **V:** How much did each individual member contribute to the final value?



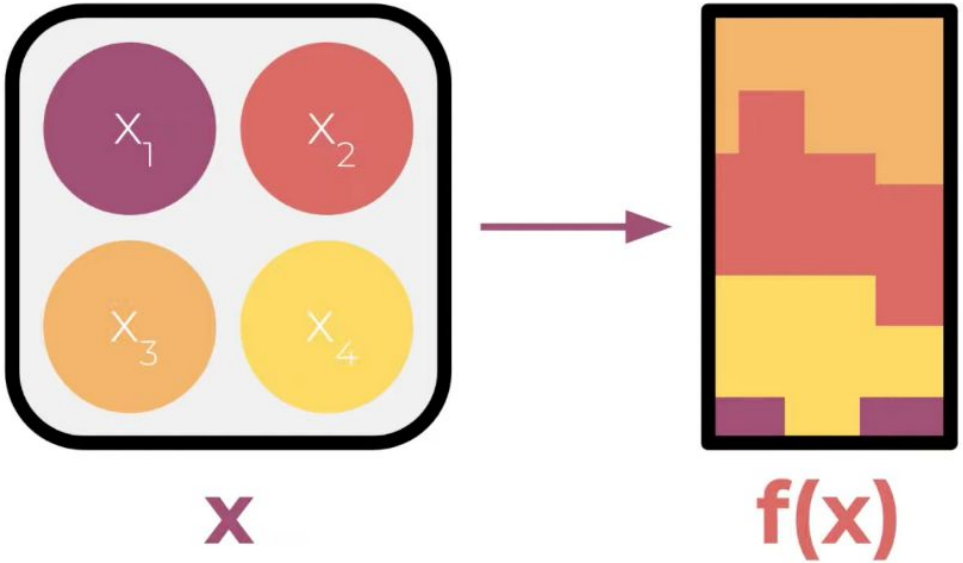
SHAP (SHapley Additive exPlanations)



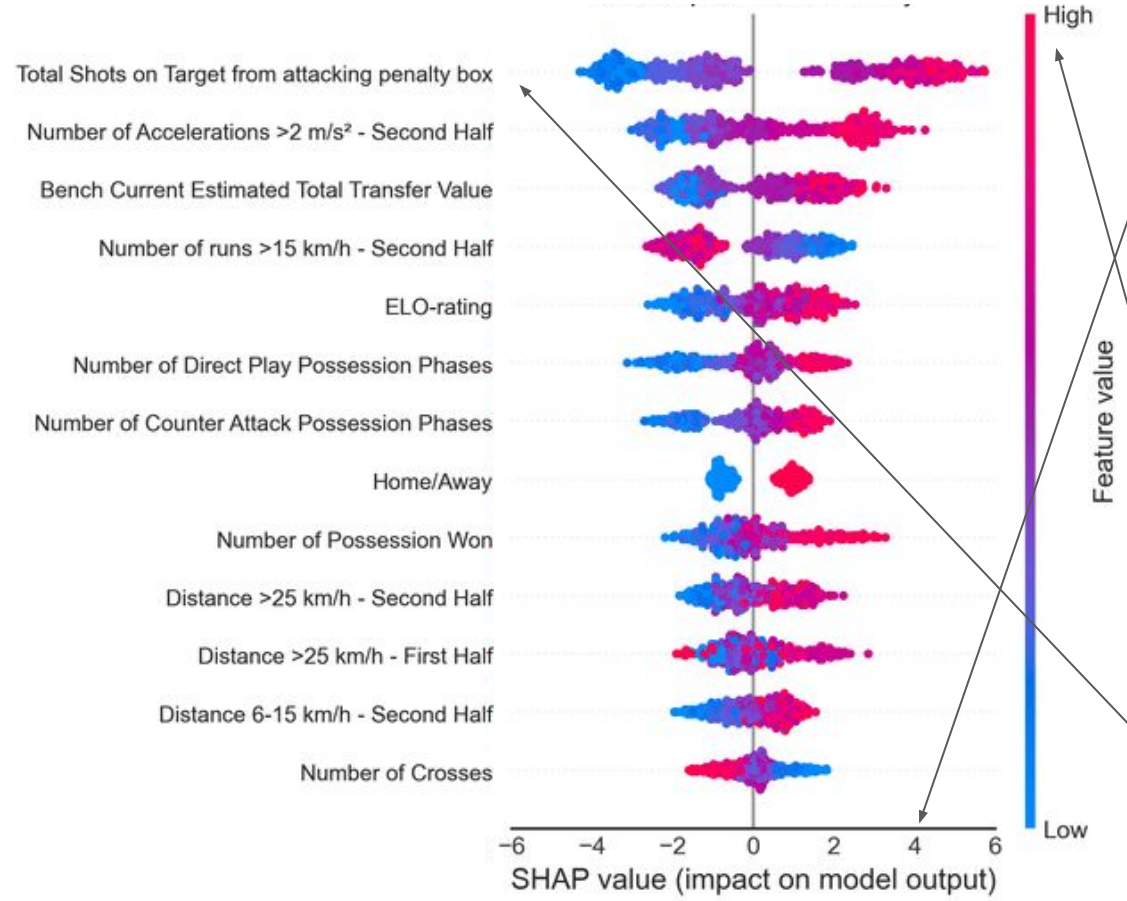
SHAP (SHapley Additive exPlanations)

$$\varphi_i = \frac{1}{\#\text{ Members}} \sum_{\forall C \text{ s.t. } i \notin C} \frac{\text{Marginal Contribution of } i \text{ to } C}{\#\text{ Coalitions of size } |C|}$$

SHAP (SHapley Additive exPlanations)



Which features contribute most to helping your team win?

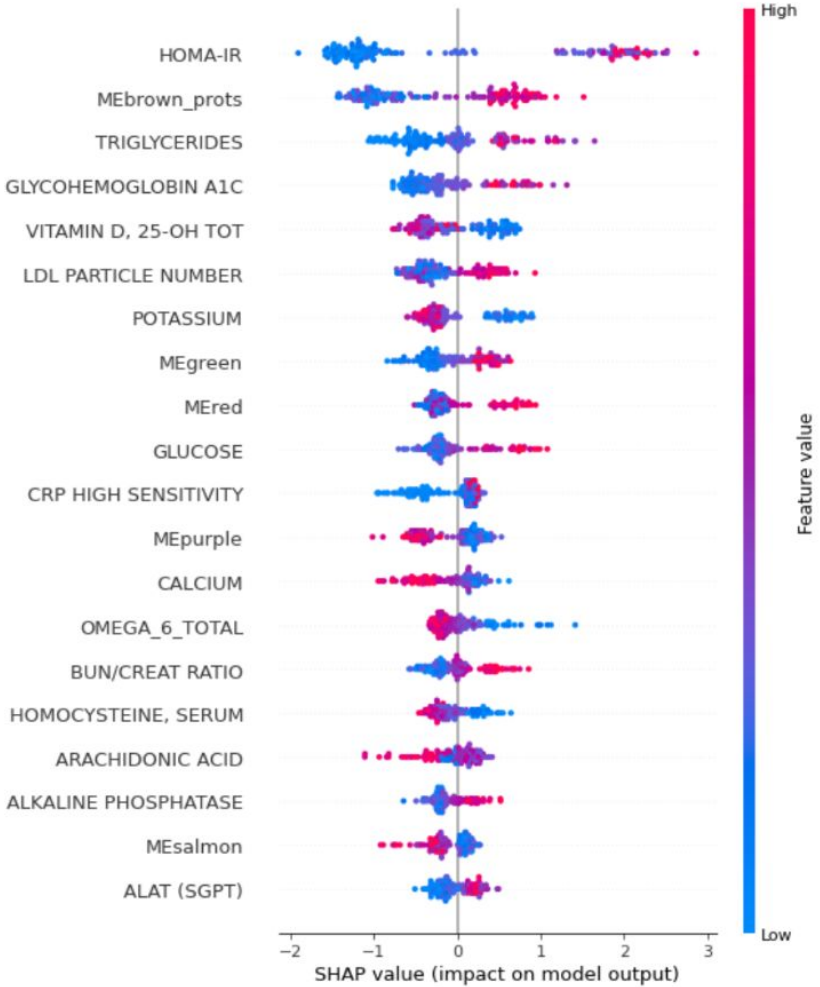


Impact shows the direction of the relationship between a feature and the game outcome

Positive SHAP values are indicative of “winning”, while negative SHAP values are indicative of “losing”

Feature Value: higher values are shown in red, while lower values are shown in blue

Order of features on plot is indicative of importance



- Metabolic features associated with frailty (high HOMA-IR, HbA1C, triglycerides). There is a historical view of frailty as low body weight, but growing evidence of a population of ‘obese frail’ with metabolic dysfunction/disease. (Mishra et al 2024)

- Low levels of Vitamin D and frailty observed previously (Zheng et al 2022), many Vitamin D supplement studies to improve frailty - mixed results.

- Low levels of calcium – linked to Vitamin D insufficiency

- Low levels of potassium – loss of muscle mass in frailty may drive hyponatremia

How could we improve our models?

- Reducing class imbalance: Exercise
- Hyperparameter tuning
- Threshold moving
- Changing our outcome definition
- Gathering more data
- Feature engineering
- Try other models