Study on Gradient Boosting Algorithms and Hyperparameter Optimization using Optuna

João Manoel Herrera Pinheiro

University of São Paulo

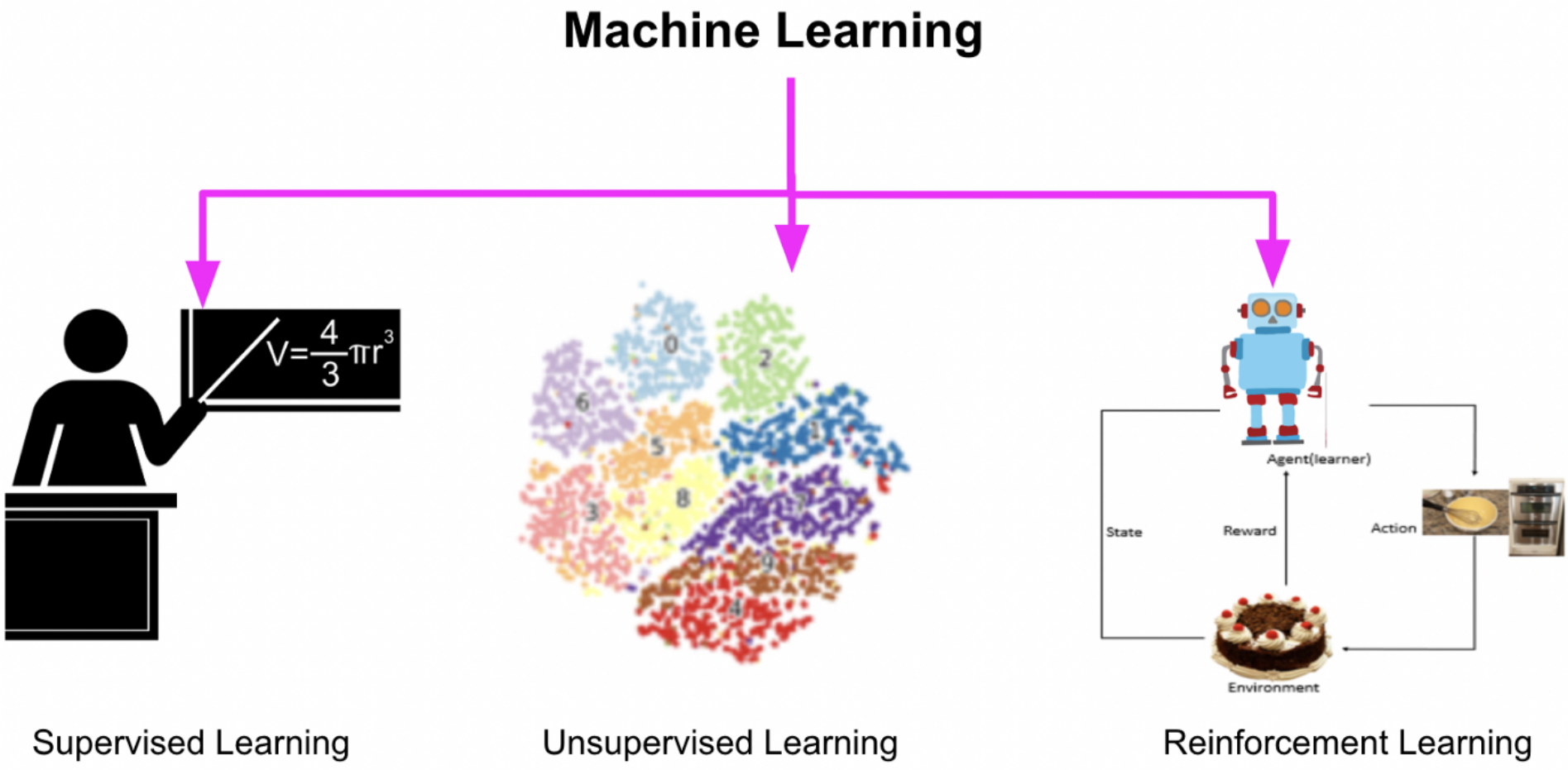
joaomh@protonmail.com

February 14, 2023

Overview

1. Introduction - 5min
   - Supervised Machine Learning
   - Machine Learning Steps
   - Binary Classification
   - Validation Metrics
   - Shapley Values
2. XGBoost, CatBoost and LightGBM - 5min
   - Gradient Boosting
   - XGBoost x CatBoost x LightGBM
3. Results: Default x Optuna - 10min
   - Disclaimer
   - Diabetes Prediction
   - Heart Failure Prediction
   - Kidney Stone Prediction
   - Breast Cancer Wisconsin Diagnostic

Supervised × Unsupervised × Reinforcement



Supervised × Unsupervised × Reinforcement

Supervised Learning

In supervised learning, the AI model is trained based on the given input and its expected output. Decision trees, linear regression, KNN, Random Forest. Image detection, Population growth prediction

Unsupervised Learning

In unsupervised learning, the AI model is trained only on the inputs, without their labels. The model classifies the input data into classes that have similar features. K-means,DBSCAN, HCA. Customer segmentation, feature elicitation, targeted marketing.

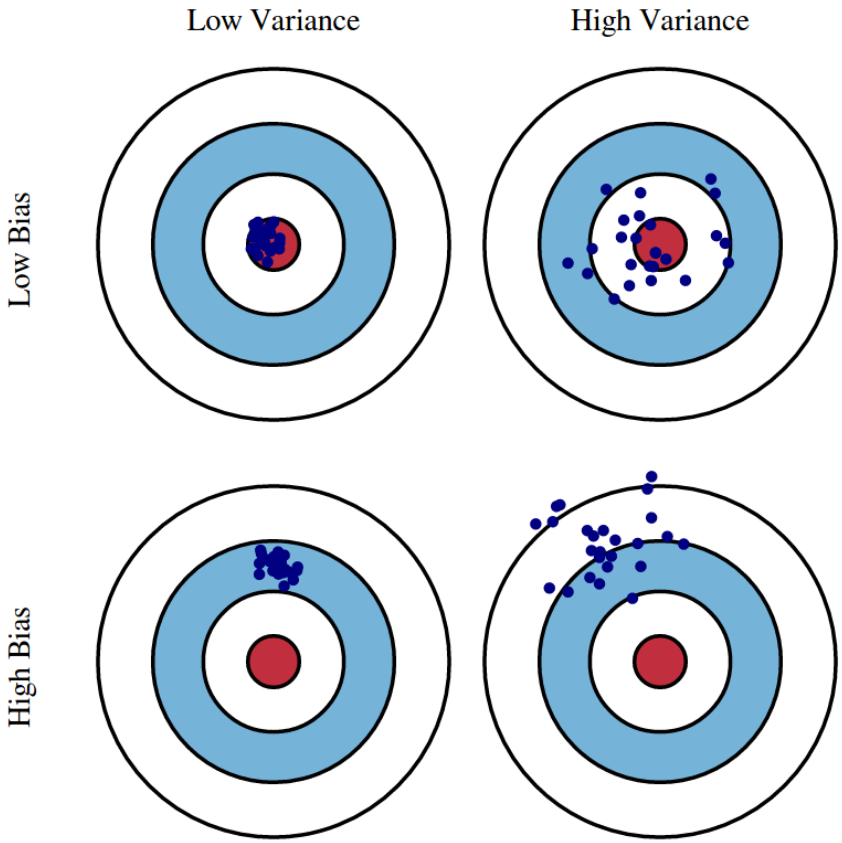
Reinforcement Learning

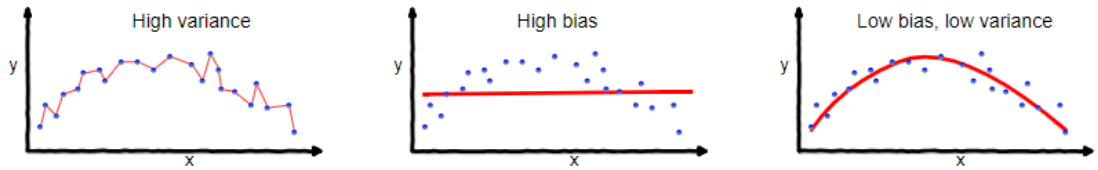
In reinforcement learning, the AI model tries to take the best possible action in a given situation to maximize the total profit. The model learns by getting feedback on its past outcomes. Deep Learning, Q-Learning, Markov decision process. Self Driving Cars.

Machine Learning Steps

- Understand the Problem
- Collecting Data
- Preparing the Data
- Choosing a Model
- Training the Model
- Evaluating the Model
- Tuning the Model
- Making Predictions

Bias–Variance Tradeoff





overfitting **underfitting** **Good balance**

Bias–Variance Tradeoff

Bias

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

Bias–Variance Tradeoff

Underfitting

Happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance.

Overfitting

Happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance.

Train-Validation-Test

Train-Validation-Test

Train

Training Dataset: The sample of data used to fit the model.The model sees and learns from this data.

Validation

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.

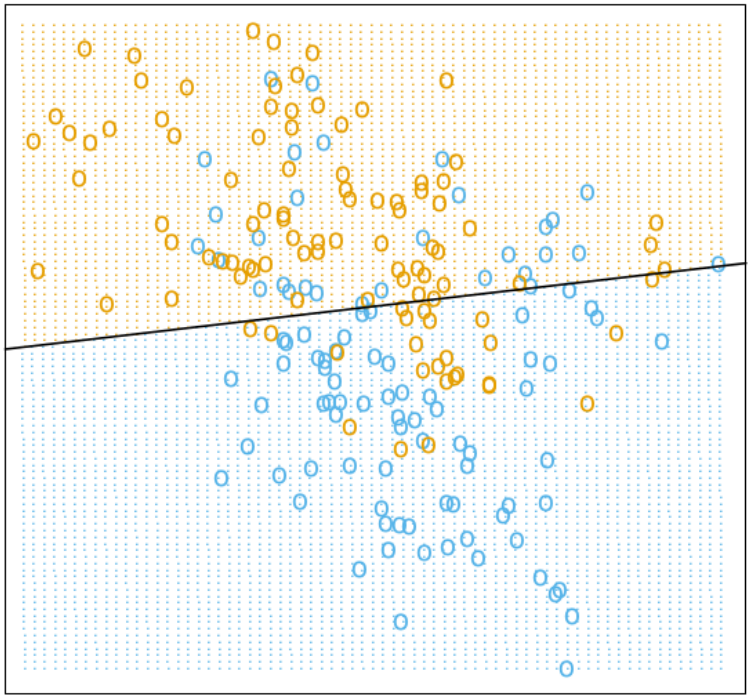
Test

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

Binary Classification

Binary Classification

Binary classification is a supervised learning algorithm that categorizes new observations into one of two classes. Often 0 or 1.



Validation Metrics:

AUC

AUC functionally measures how well-ordered results are in accordance with true class membership. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

Logloss

Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.

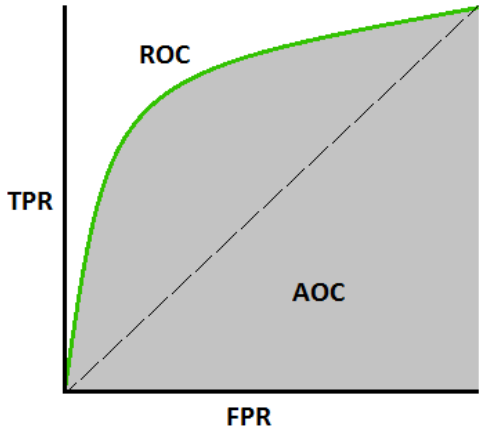
KS

The KS statistic for two samples is simply the highest distance between their two CDFs, so if we measure the distance between the positive and negative class distributions, we can have another metric to evaluate classifiers.

Validation Metrics: AUC

ROC

The ROC curve is plotted with the True Positive Rate, or Sensitivity, against False Positive Rate, or 1 - Specificity.



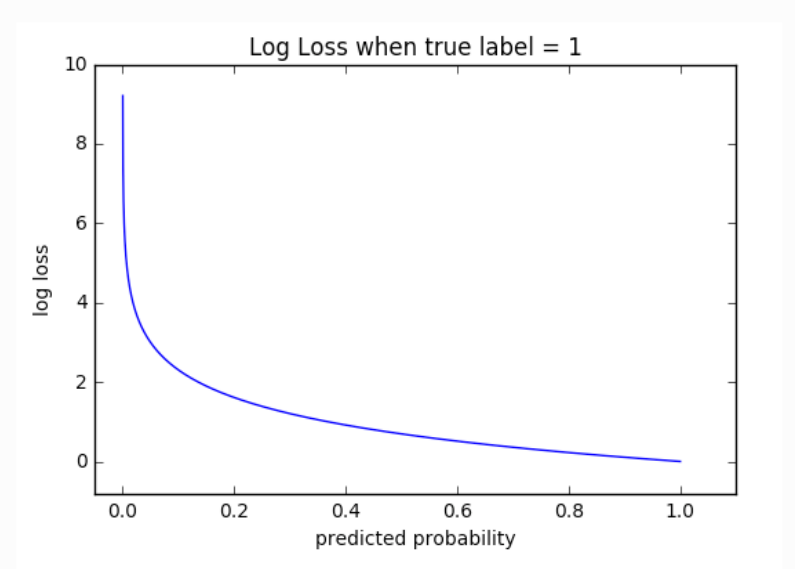
$$Roc_y(\mathcal{T}) = Sensitivity = TPR_{\mathcal{T}} = \frac{TP_{\mathcal{T}}}{TP_{\mathcal{T}} + FN_{\mathcal{T}}}$$

$$Roc_x(\mathcal{T}) = 1 - Specificity = FPR_{\mathcal{T}} = \frac{FP_{\mathcal{T}}}{FP_{\mathcal{T}} + TN_{\mathcal{T}}}$$

Validation Metrics: Logloss

Logloss

Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The more the predicted probability diverges from the actual value, the higher is the log-loss value.

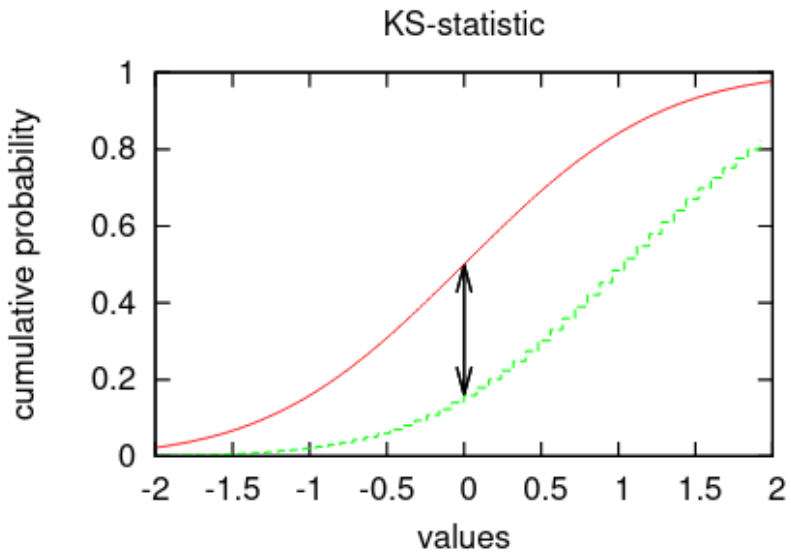


$$Logloss = -\frac{1}{N}\sum_{i=1}^{n}[y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})]$$

Validation Metrics: Kolmogorov-Smirnov Test

KS

The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples.



$$KS = \max(TPR - FPR)$$

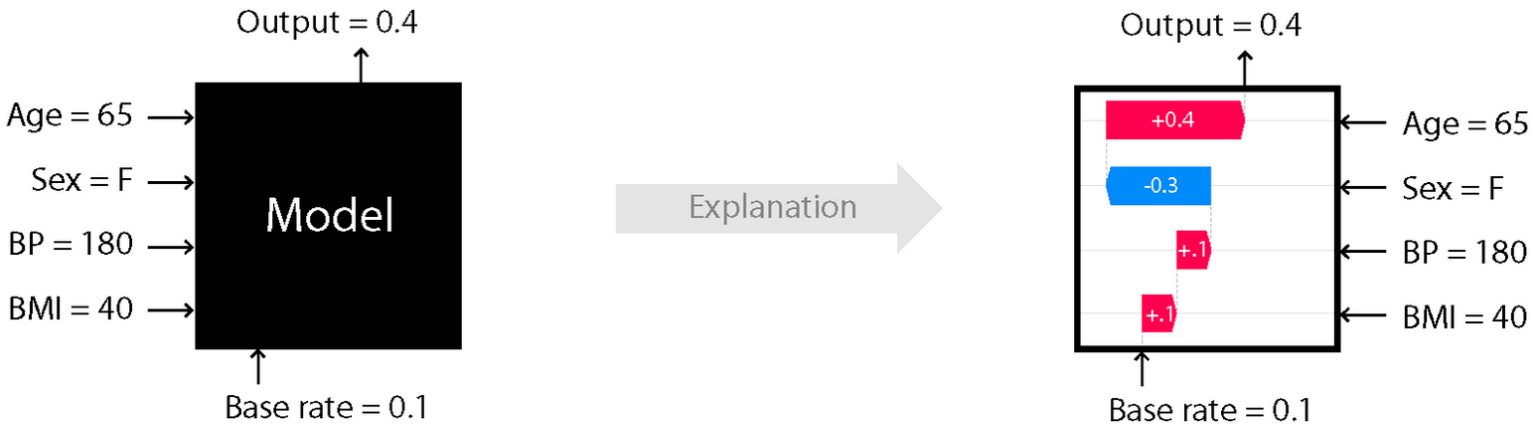
Shapley Values

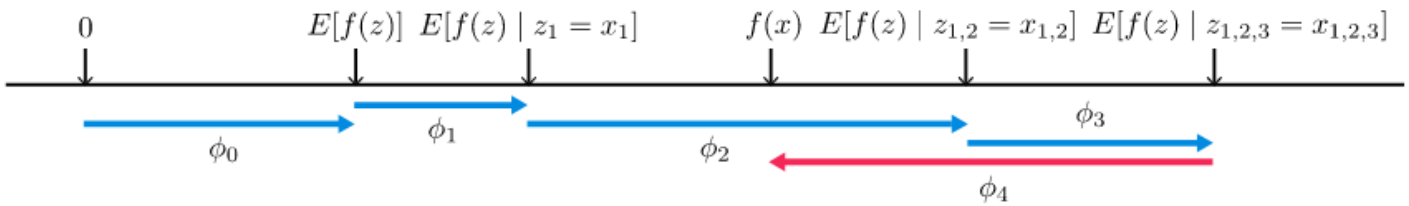
SHAP (SHapley Additive exPlanations)

SHAP is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i$$

SHAP

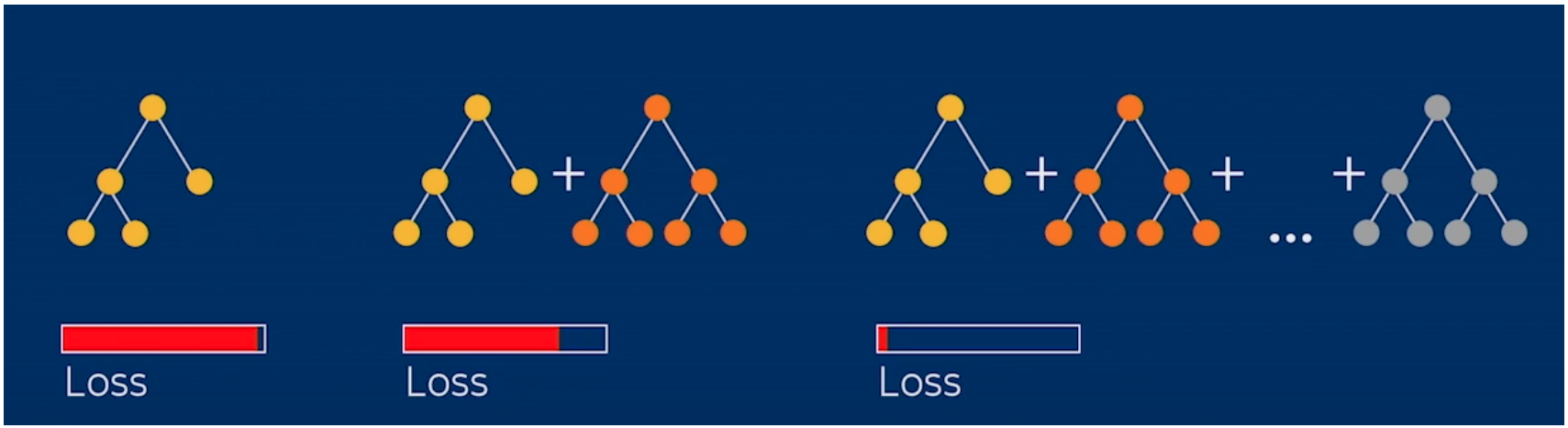

Output = 0.4
Age = 65
Sex = F
BP = 180
BMI = 40
Model
Base rate = 0.1
Explanation
Output = 0.4
+0.4
-0.3
+.1
+.1
Age = 65
Sex = F
BP = 180
BMI = 40
Base rate = 0.1


0
$E[f(z)]$
$E[f(z) \mid z_1 = x_1]$
$f(x)$
$E[f(z) \mid z_{1,2} = x_{1,2}]$
$E[f(z) \mid z_{1,2,3} = x_{1,2,3}]$
ϕ_0
ϕ_1
ϕ_2
ϕ_3
ϕ_4

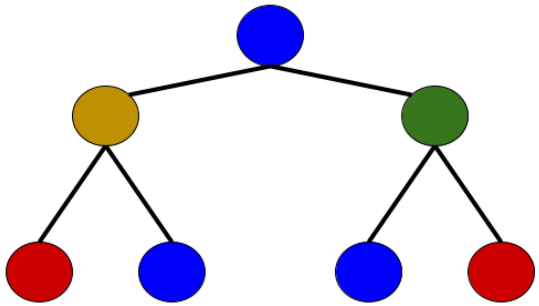
Gradient Boosting

Boosting

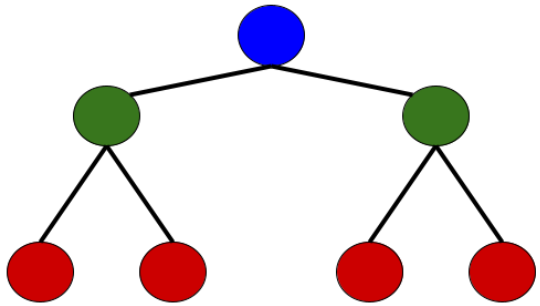
Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set.



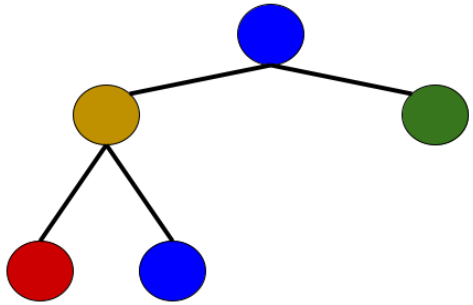
XGBoost x CatBoost x LightGBM



XGBoost



CatBoost



LightGBM

XGBoost x CatBoost x LightGBM

	CatBoost	LightGBM	XGBoost
Developer	Yandex	Microsoft	DMLC
Release Year	2017	2016	2014
Tree Symmetry	Symmetric	Asymmetric Leaf-wise tree growth	Asymmetric Level-wise tree growth
Splitting Method	Greedy method	Gradient-based One-Side Sampling (GOSS)	Pre-sorted and histogram-based algorithm
Type of Boosting	Ordered	-	-
Numerical Columns	Support	Support	Support
Categorical Columns	Support Perform one-hot encoding (default) Transforming categorical to numerical columns by border, bucket, binarized target mean value, counter methods available	Support, but must use numerical columns Can interpret ordinal category	Support, but must use numerical columns Cannot interpret ordinal category, users must convert to one-hot encoding, label encoding or mean encoding
Text Columns	Support Support Bag-of-Words, Naïve-Bayes or BM-25 to calculate numerical features from text data	Do not support	Do not support
Missing values	Handle missing value Interpret as NaN (default) Possible to interpret as error, or processed as minimum or maximum values	Handle missing value Interpret as NaN (default) or zero Assign missing values to side that reduces loss the most in each split	Handle missing value Interpret as NaN (tree booster) or zero (linear booster) Assign missing values to side that reduces loss the most in each split

Hyperparameter Space

learning_rate , max_depth, reg_lambda

$$LR = \{0.1, 0.3\}$$
$$MD = \{3, 6\}$$
$$RegL = \{0.01, 0.05\}$$

Hyperparameter Space

	XGBoost	CatBoost	LightGBM
AUC	0.92598	0.89428	0.89406
logloss	2.17379	3.13992	2.89838
KS	0.85195	0.78856	0.78812
time (s)	0.10825	1.05492	0.09304

Table: $LR = 0.1, MD = 3, Reg_L = 0.01$

	XGBoost	CatBoost	LightGBM
AUC	0.92598	0.90448	0.92598
logloss	2.17379	2.41531	2.17379
KS	0.85195	0.80895	0.85195
time (s)	0.11743	1.00966	0.08180

Table: $LR = 0.3, MD = 3, Reg_L = 0.01$

Diabetes Prediction

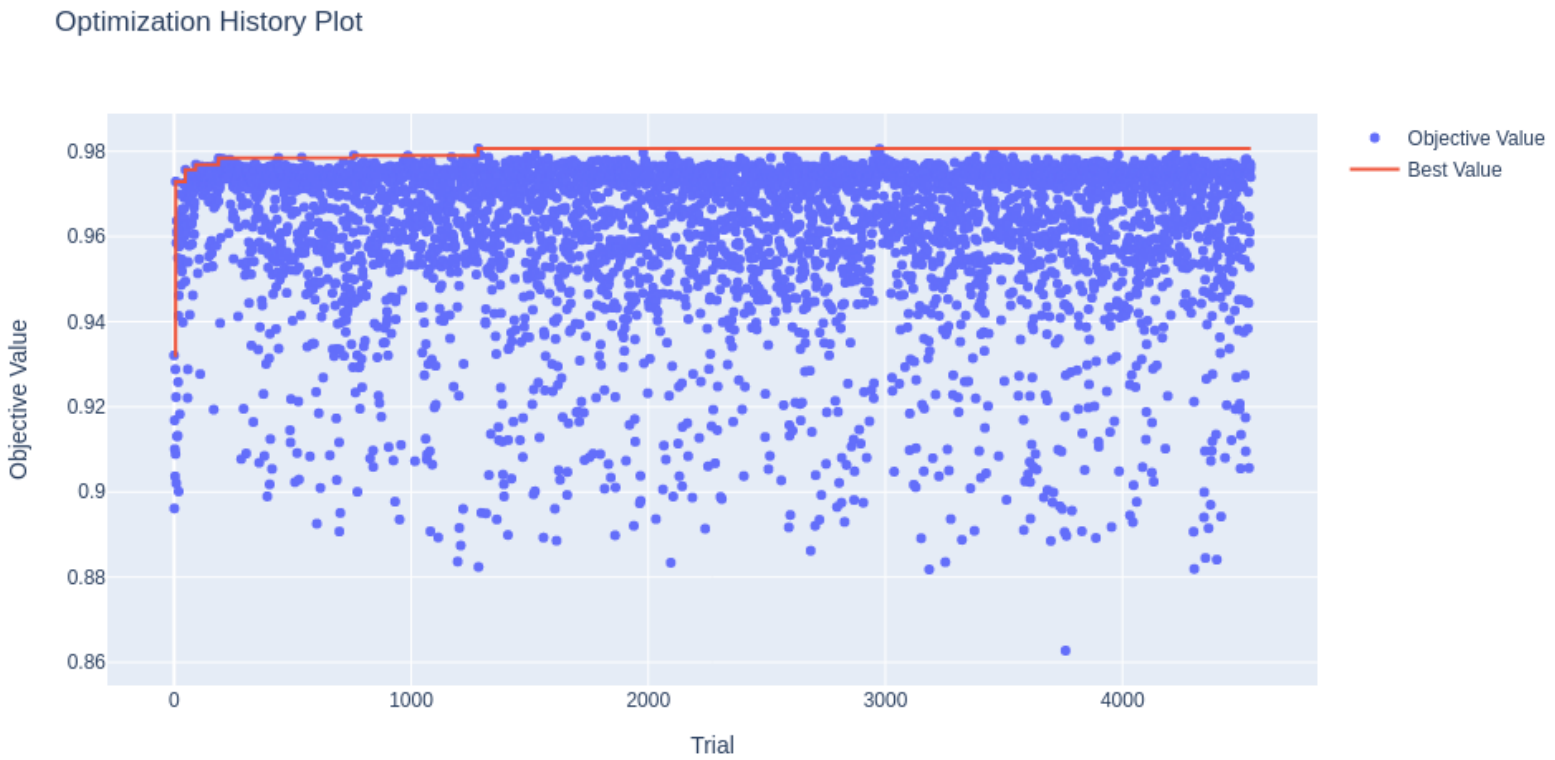
	XGBoost	CatBoost	LightGBM
AUC	0.93639	0.89655	0.92066
logloss	1.69072	2.65684	2.29456
KS	0.87278	0.79311	0.84131
time (s)	0.17799	1.78698	0.08261

Table: Default

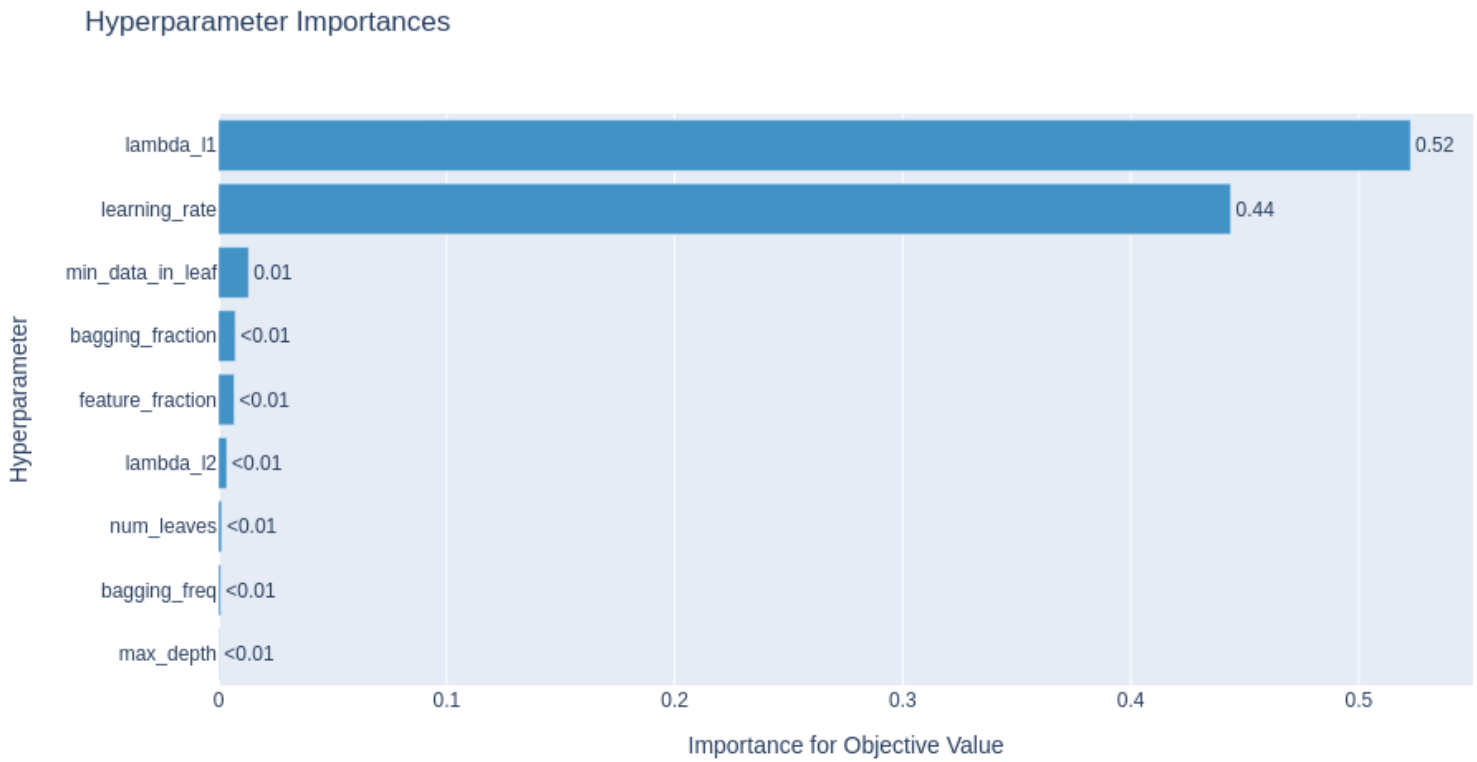
	XGBoost	CatBoost	LightGBM
AUC	0.97595	0.981106	0.98166
logloss	0.17300	0.17327	0.18599
KS	0.87821	0.87799	0.88320
δ_{AUC}	4.22%	9.43%	6.52%
δ_{KS}	0.62	10.70%	4.98%

Table: Optuna

Diabetes: LightGBM

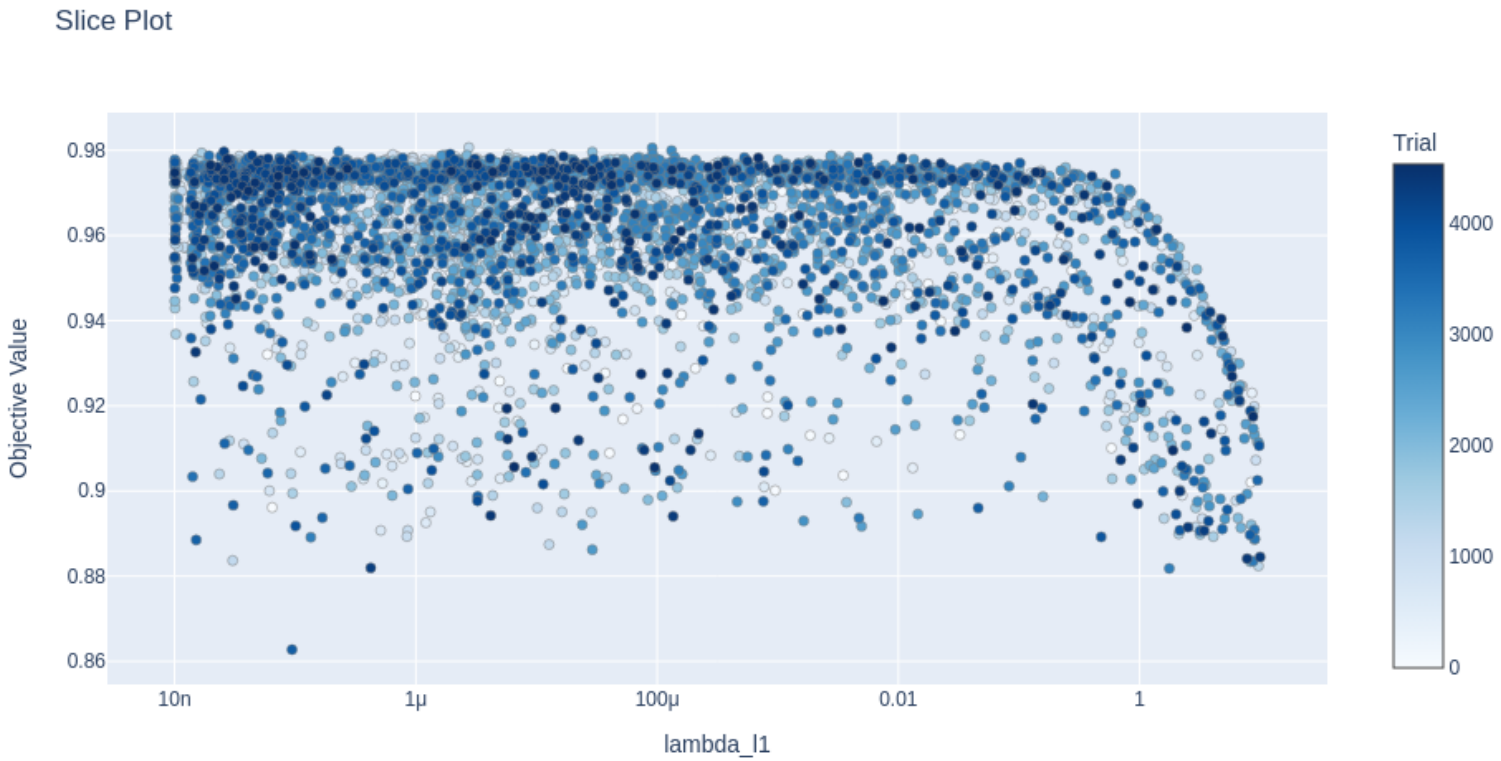


Diabetes: LightGBM



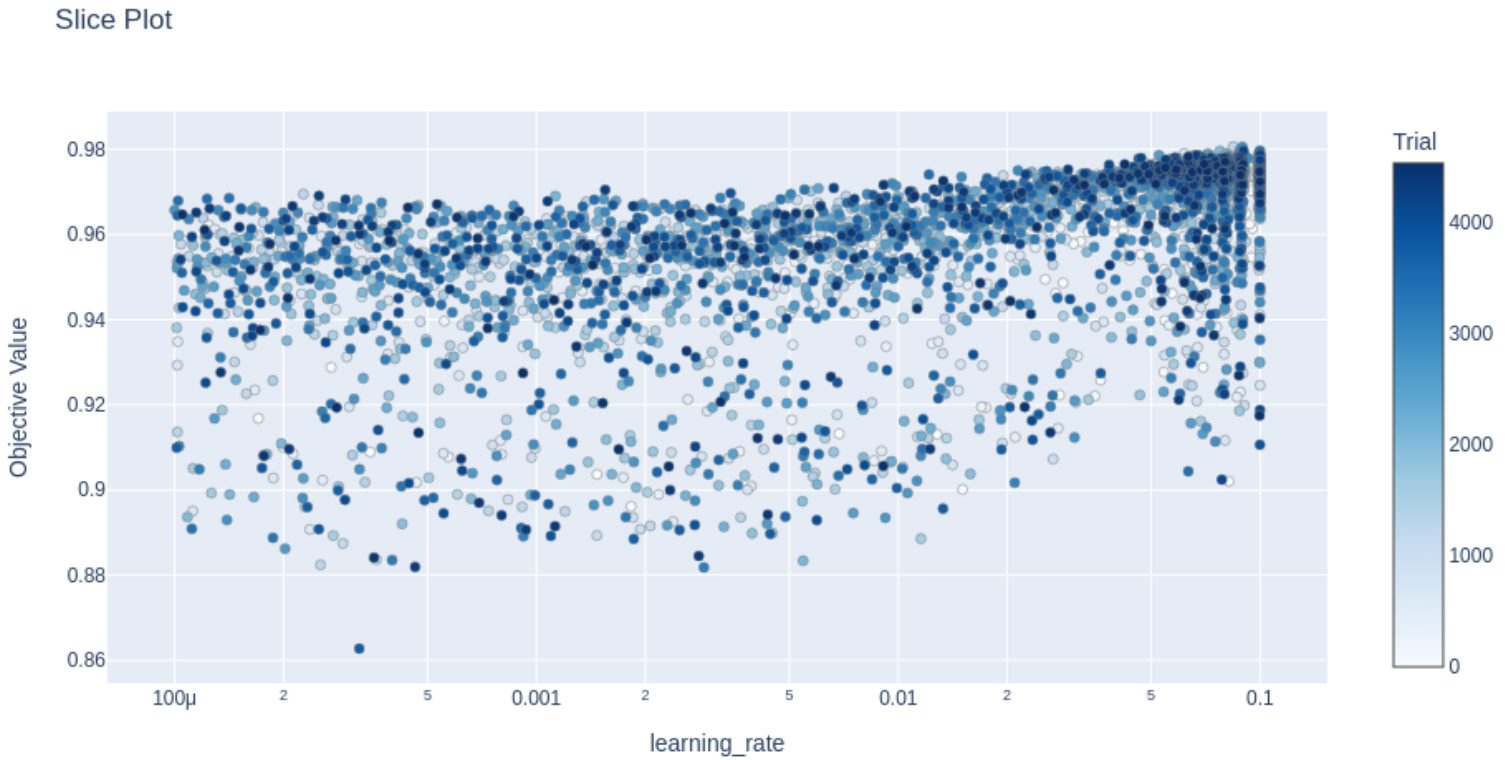
Diabetes: LightGBM

'lambda_l1': 2.7481689793447196e-06



Diabetes: LightGBM

'learning_rate': 0.08425779644832665



Diabetes: SHAP LightGBM

Heart Failure Prediction

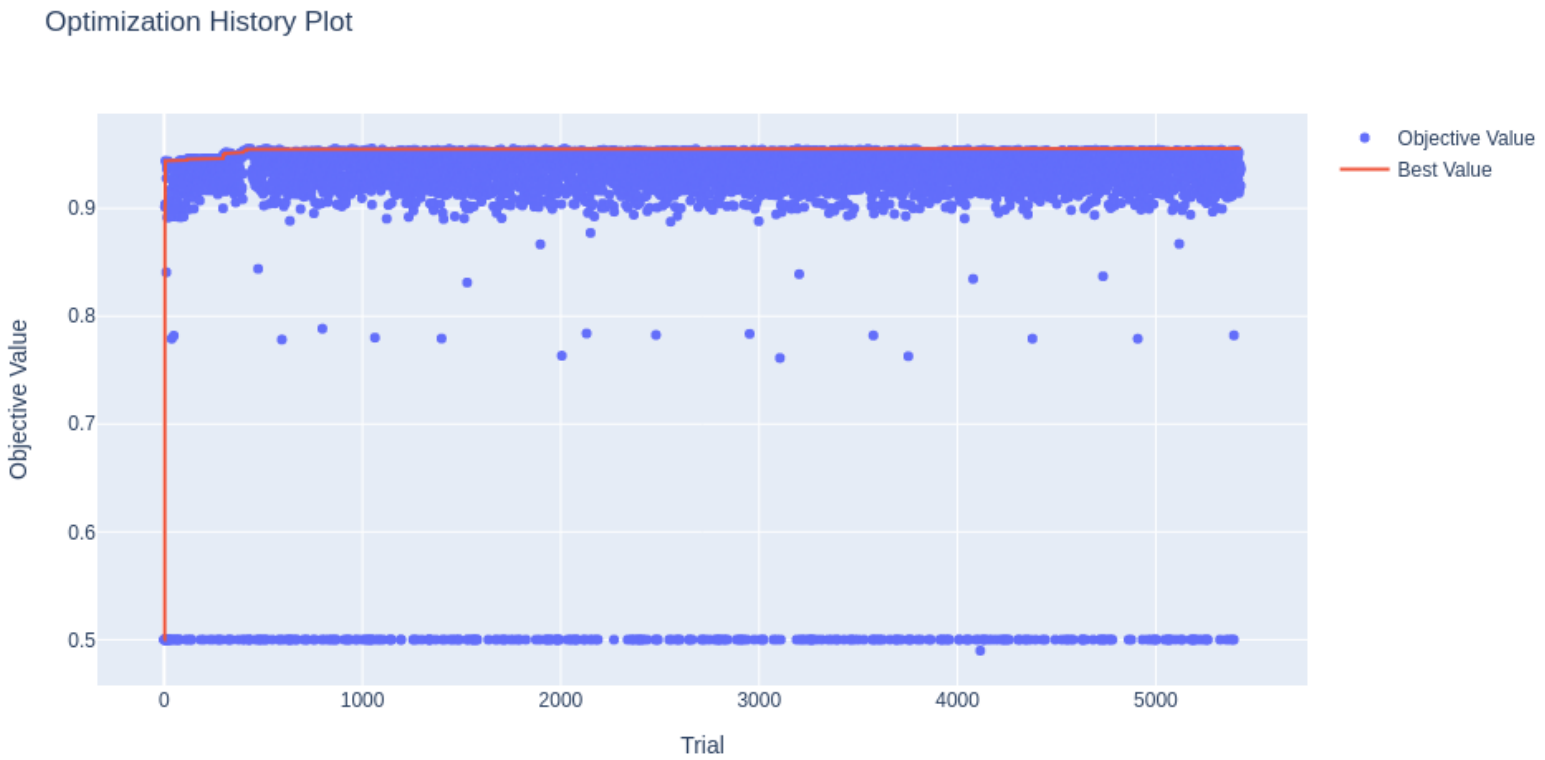
	XGBoost	CatBoost	LightGBM
AUC	0.84930	0.90048	0.86291
logloss	5.25595	3.50396	0.75538
KS	0.69861	0.80096	0.72583
time (s)	0.13467	1.41323	0.06788

Table: Default

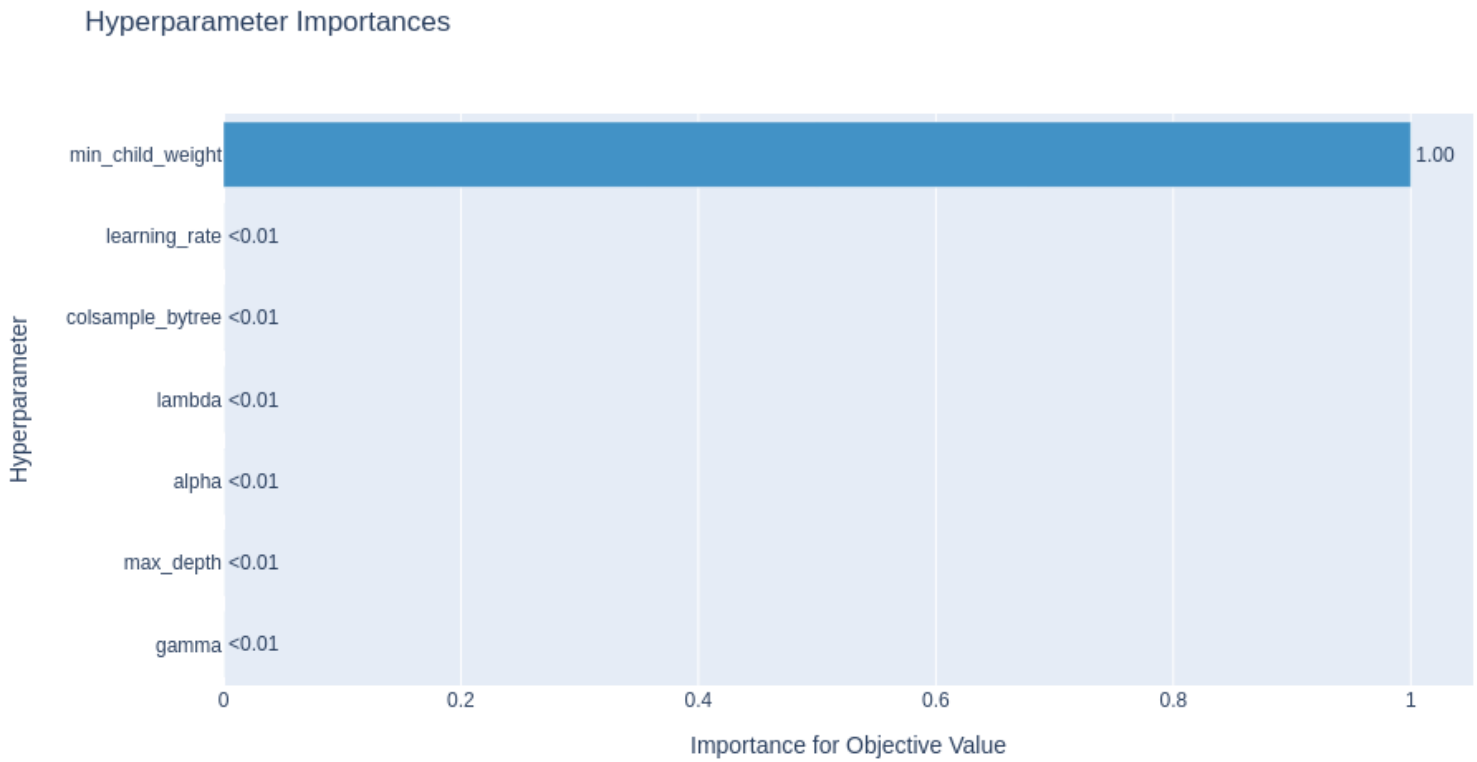
	XGBoost	CatBoost	LightGBM
AUC	0.95852	0.95019	0.95835
logloss	0.28247	0.30123	0.29087
KS	0.81598	0.78680	0.80945
δ_{AUC}	12.51%	5.52%	11.06%
δ_{KS}	16.80%	-1.77%	11.52%

Table: Optuna

Heart Failure Prediction: XGBoost

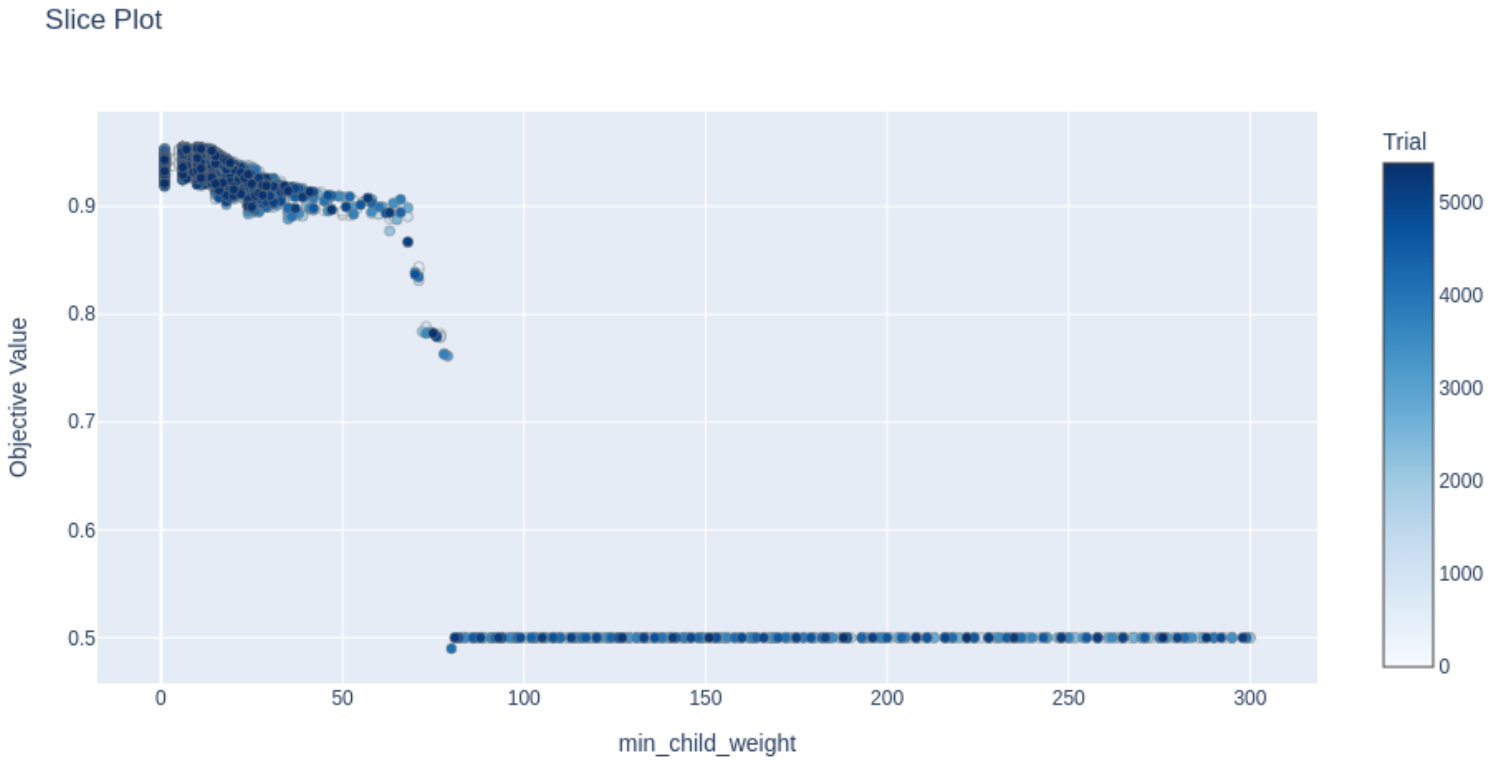


Heart Failure Prediction: XGBoost

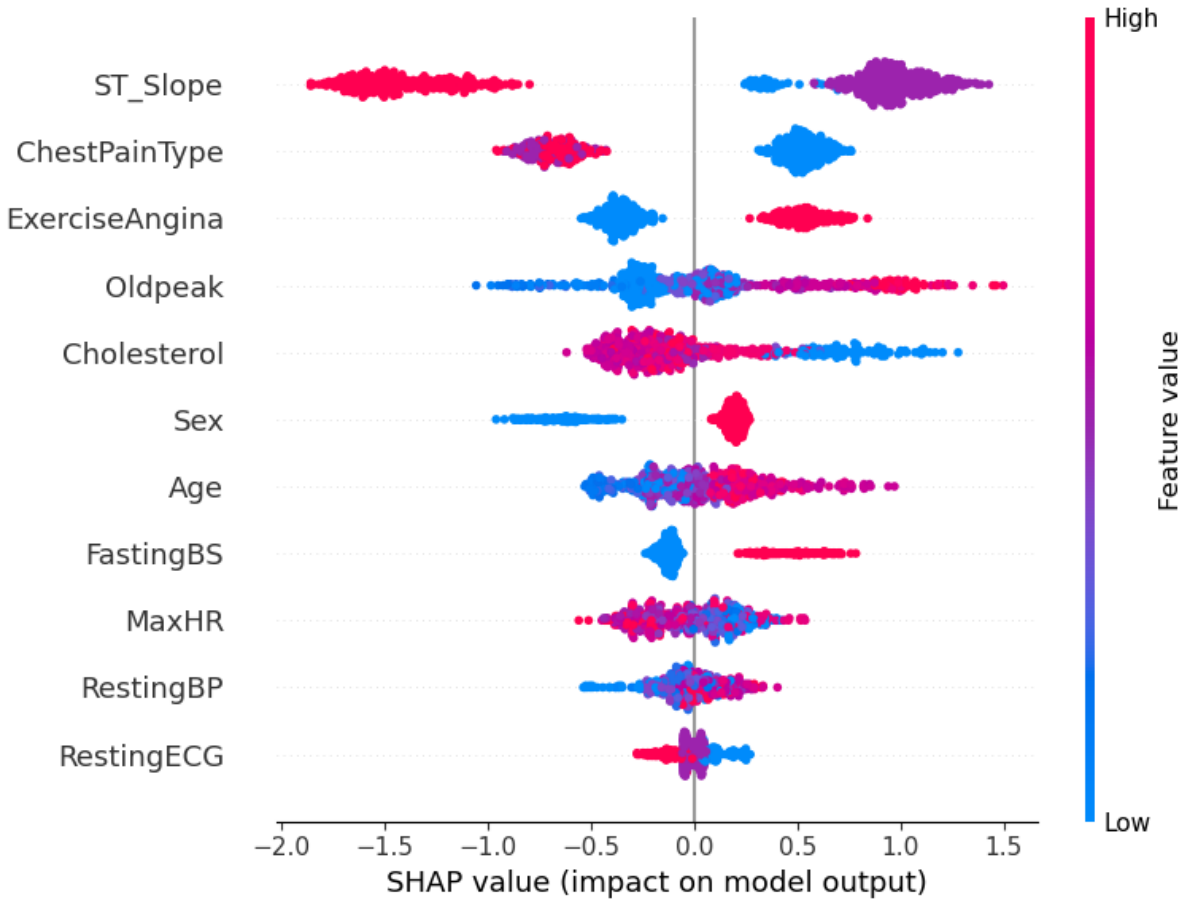


Heart Failure Prediction: XGBoost

'min_child_weight': 6



Heart Failure Prediction: SHAP XGBoost



Kidney Stone Prediction

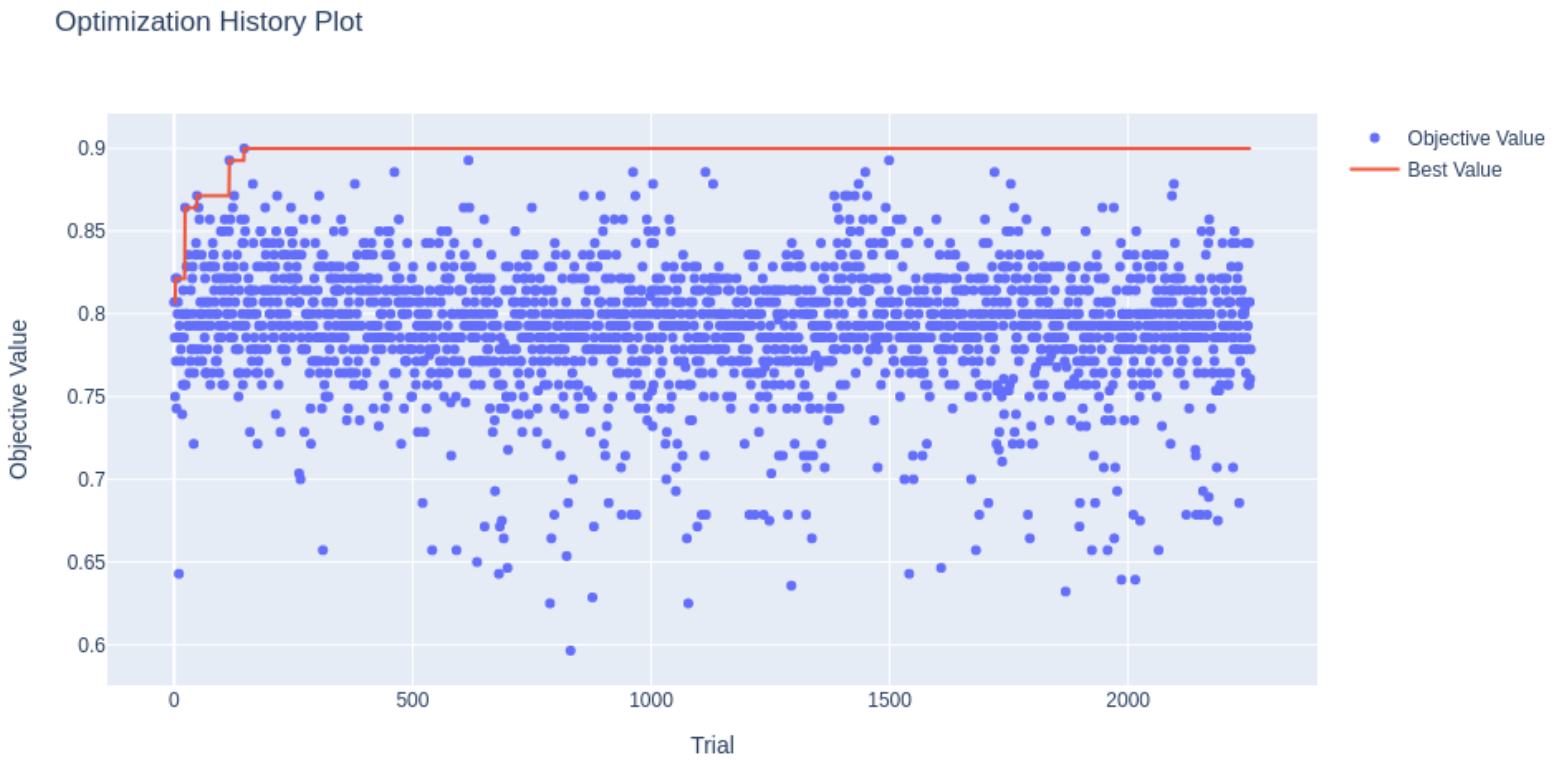
	XGBoost	CatBoost	LightGBM
AUC	0.67857	0.74286	0.72857
logloss	10.07388	8.63479	8.63476
KS	0.35714	0.48571	0.45714
time (s)	0.10748	0.27430	0.09398

Table: Default

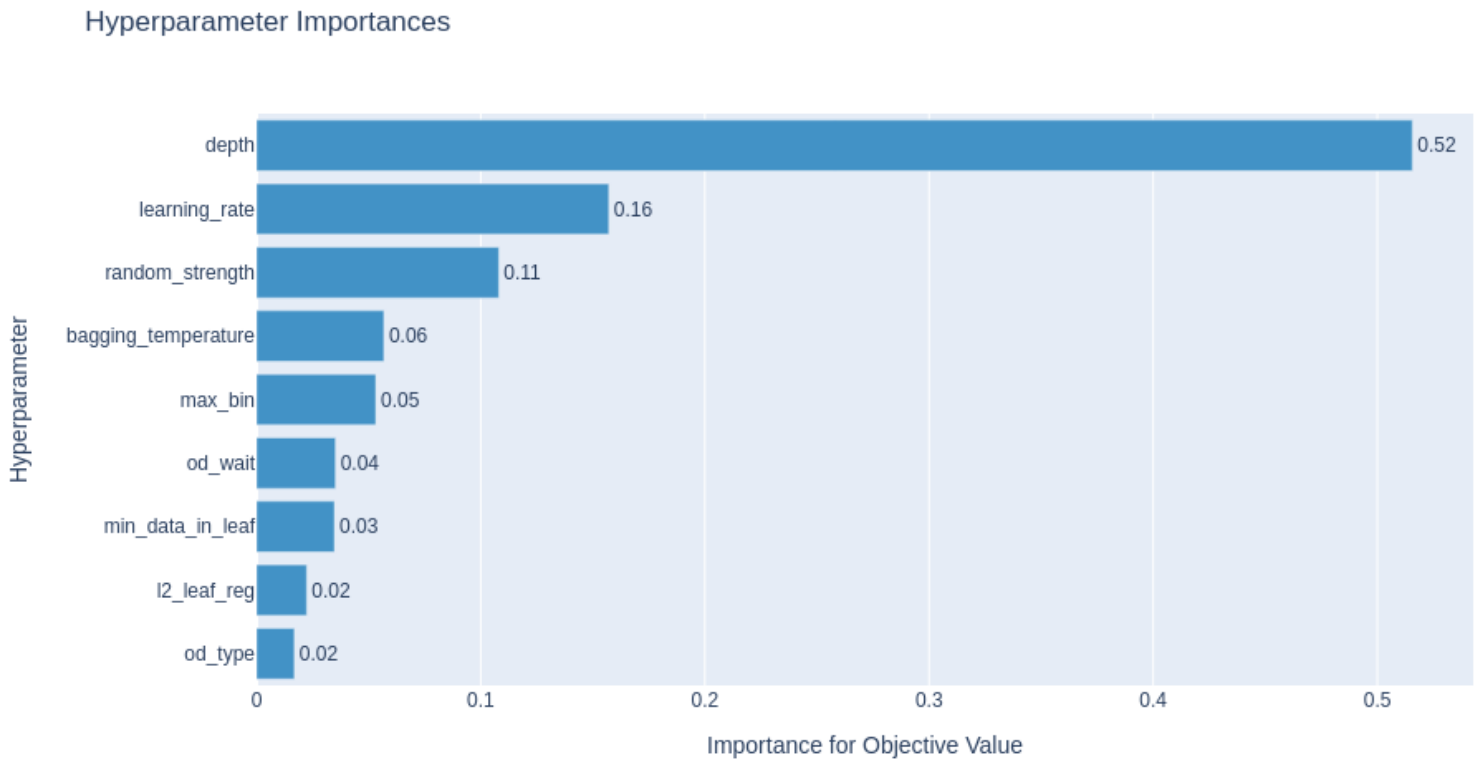
	XGBoost	CatBoost	LightGBM
AUC	0.80000	0.90000	0.88571
logloss	0.57024	0.96512	0.43693
KS	0.51429	0.71429	0.72857
δ_{AUC}	31.76%	23.53%	21.57%
δ_{KS}	140.00%	56.26%	59.37%

Table: Optuna

Kidney Stone Prediction: CatBoost

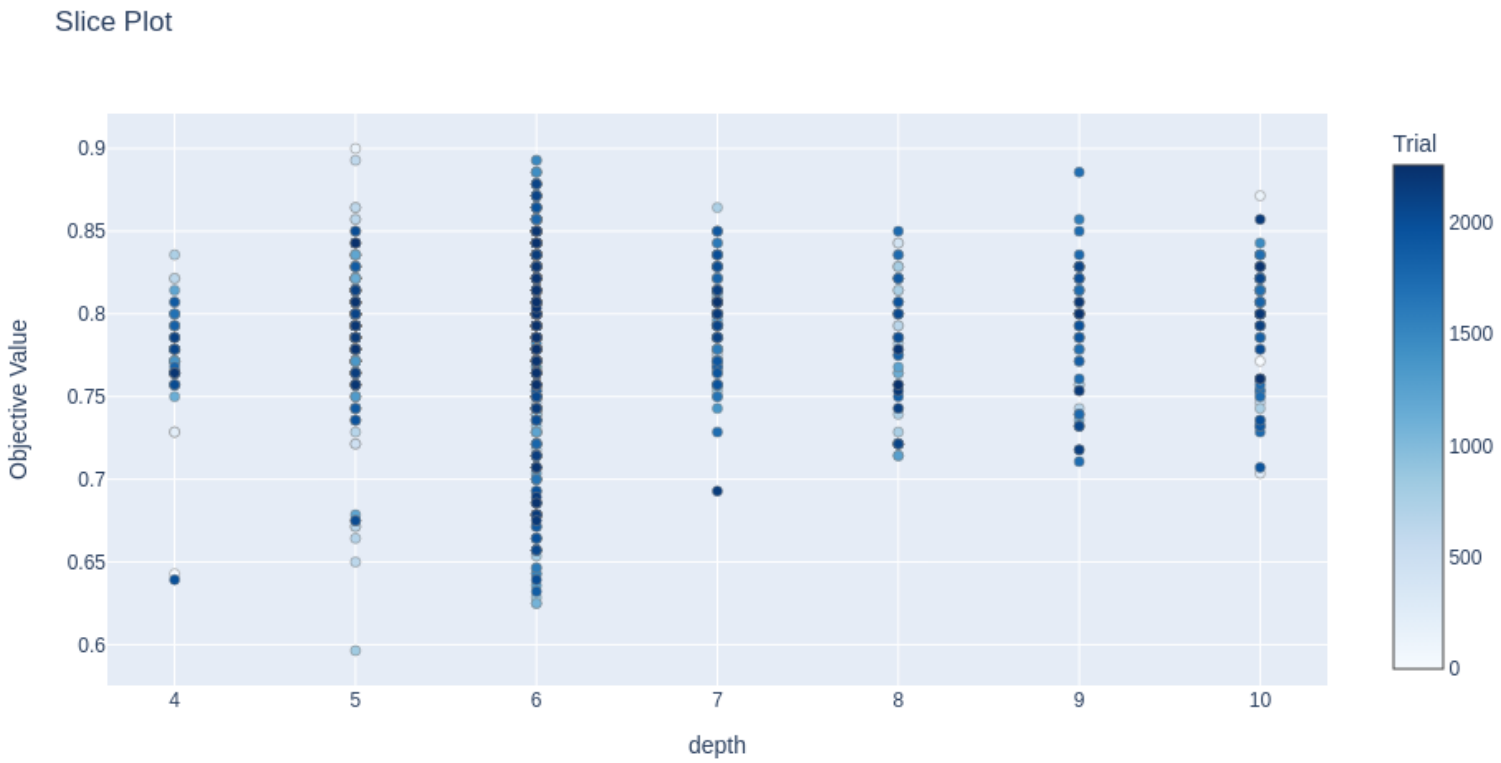


Kidney Stone Prediction: CatBoost



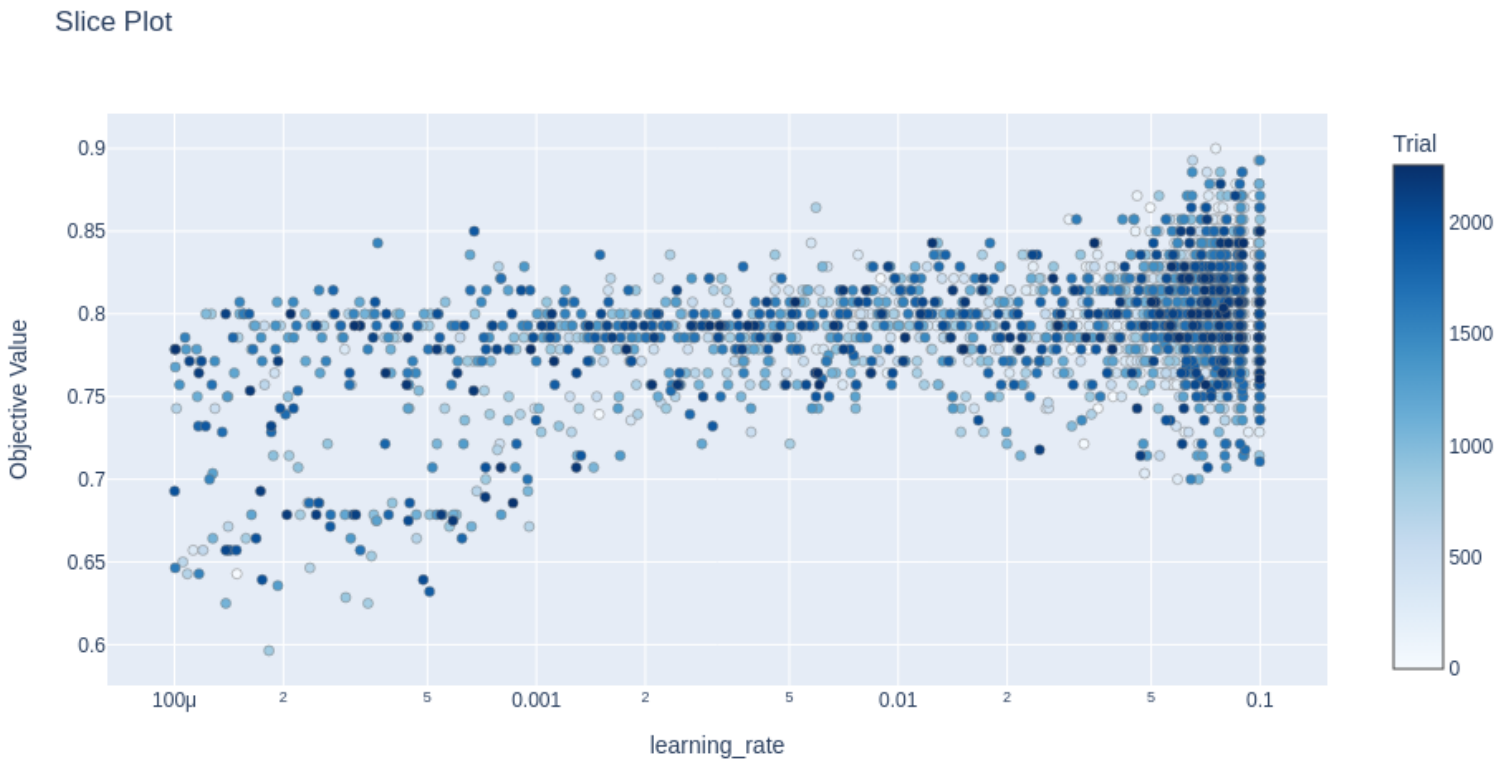
Kidney Stone Prediction: CatBoost

'depth': 5

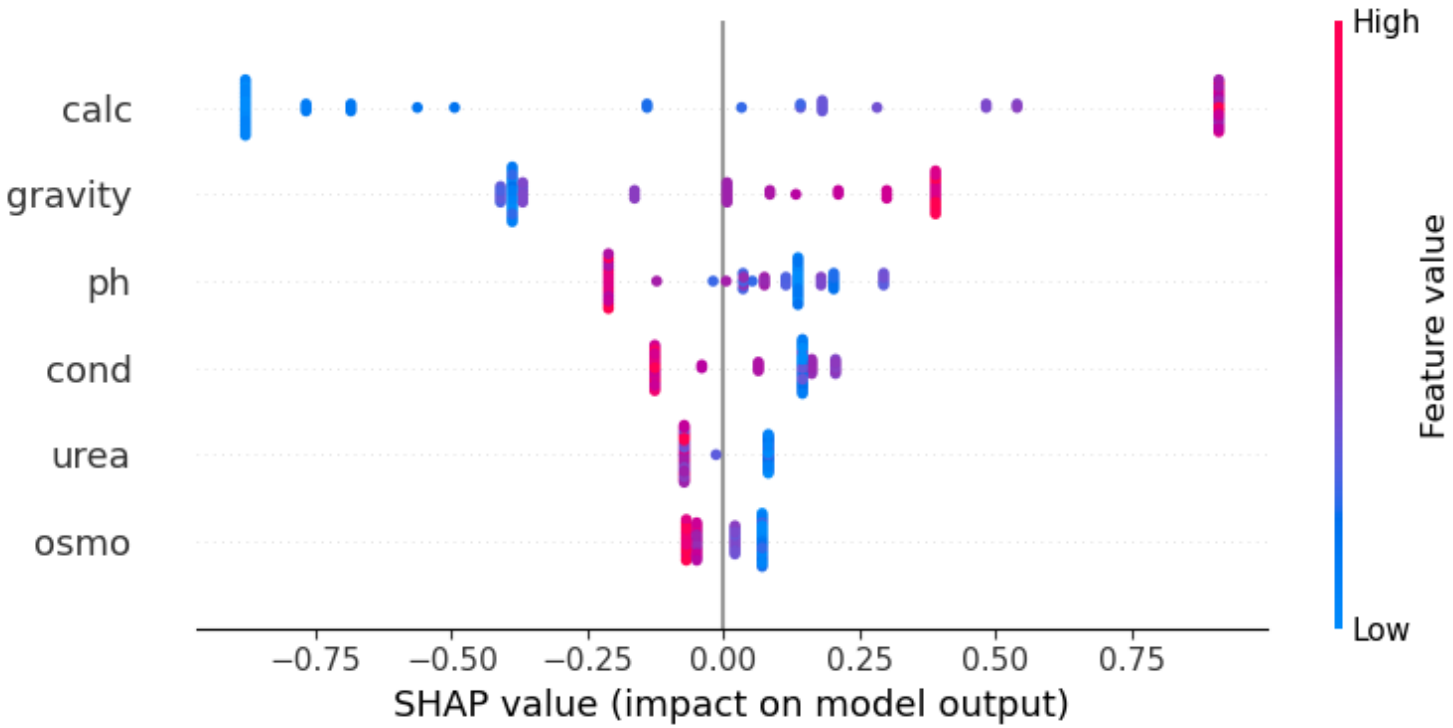


Kidney Stone Prediction: CatBoost

'learning_rate': 0.07537894328903638



Kidney Stone Prediction: SHAP CatBoost



Breast Cancer Wisconsin: XGBoost

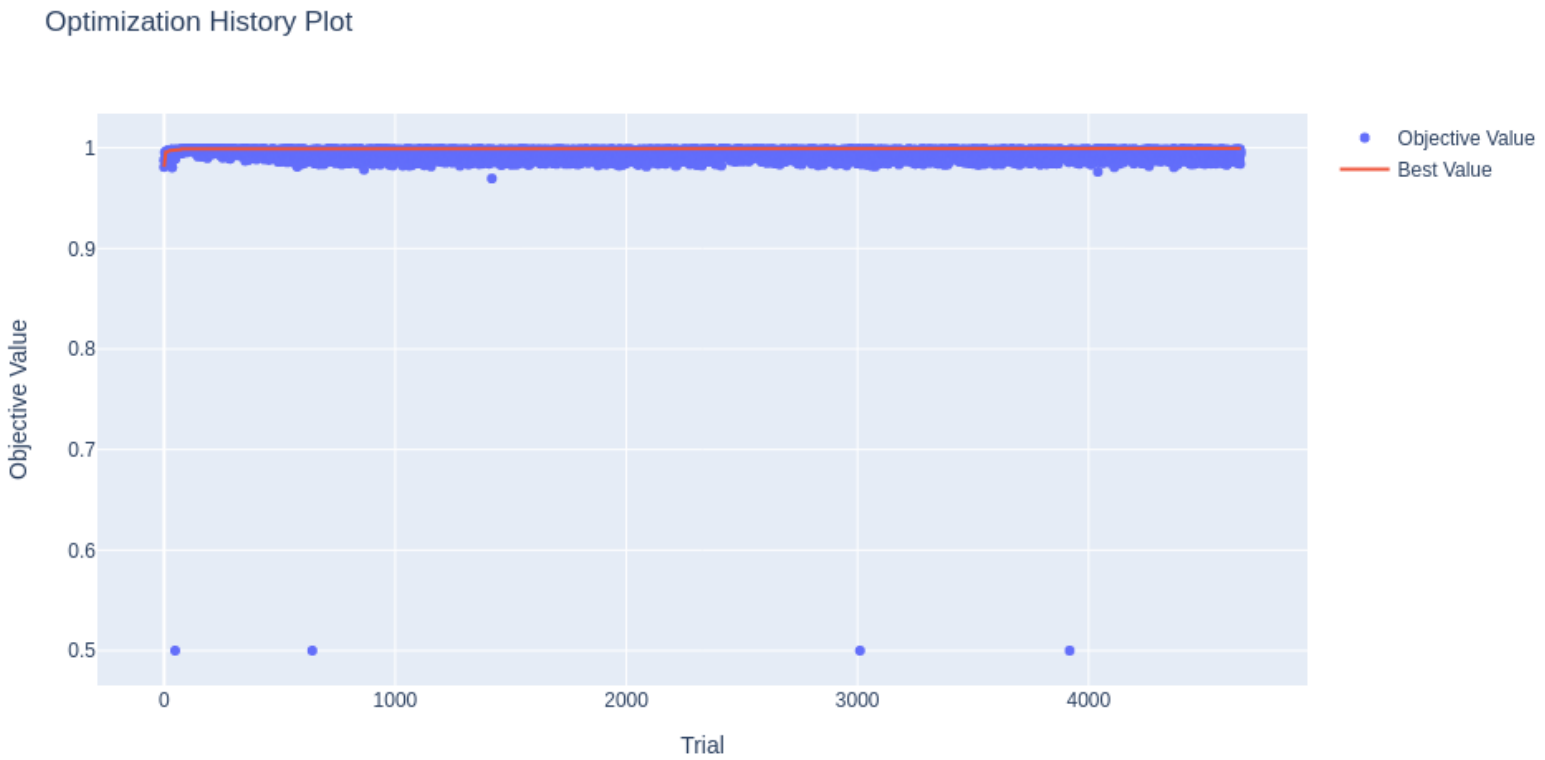
	XGBoost	**CatBoost**	**LightGBM**
AUC	0.97950	0.97156	0.94511
logloss	0.60595	0.80793	1.81785
KS	0.95899	0.94312	0.89021
time (s)	0.14904	2.72823	0.08682

Table: Default

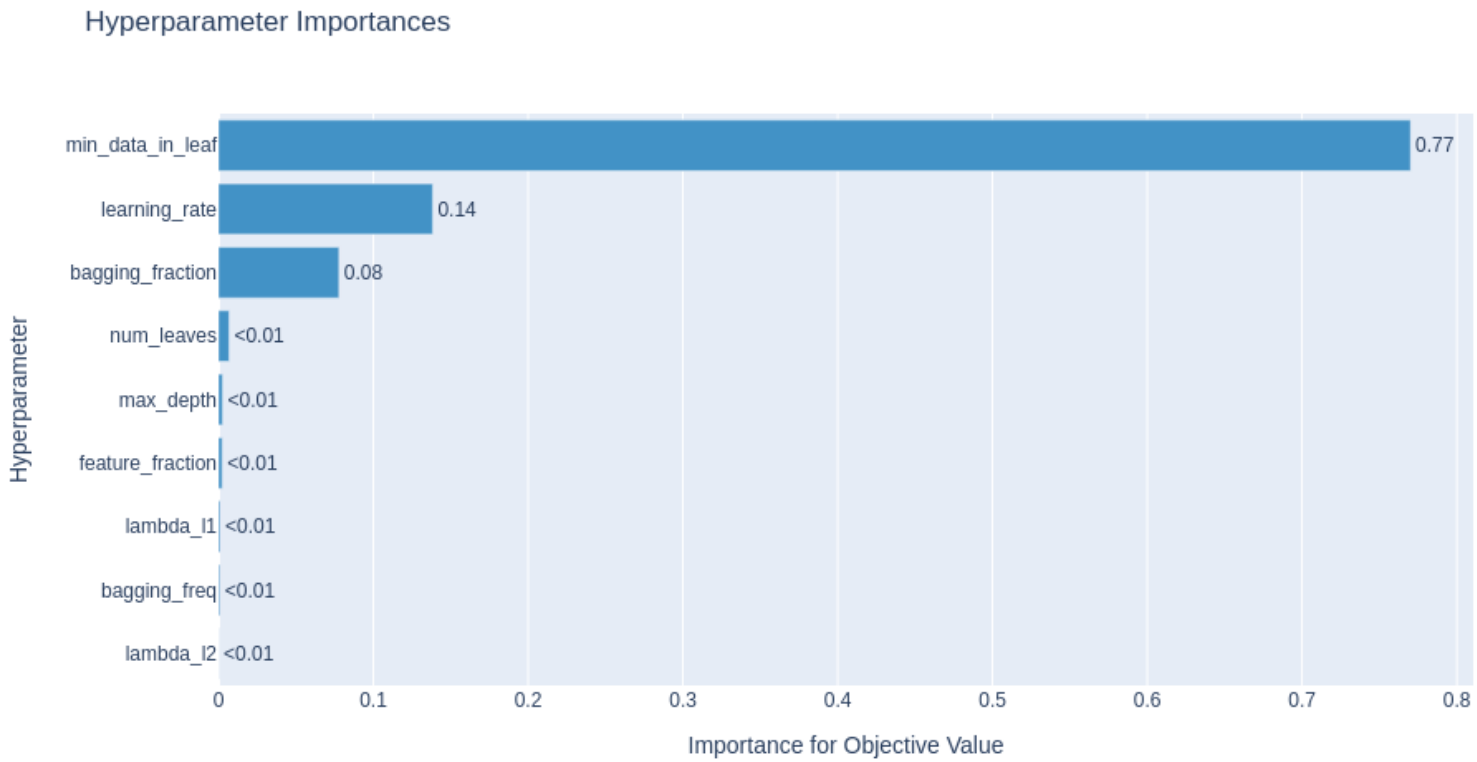
	XGBoost	**CatBoost**	**LightGBM**
AUC	0.99882	0.99927	0.99941
logloss	0.11639	0.04454	0.06708
KS	0.98413	0.97487	0.98413
δ_{AUC}	1.97%	2.85%	5.75%
δ_{KS}	2.62%	3.37%	10.55%

Table: Optuna

Breast Cancer Wisconsin: LightGBM

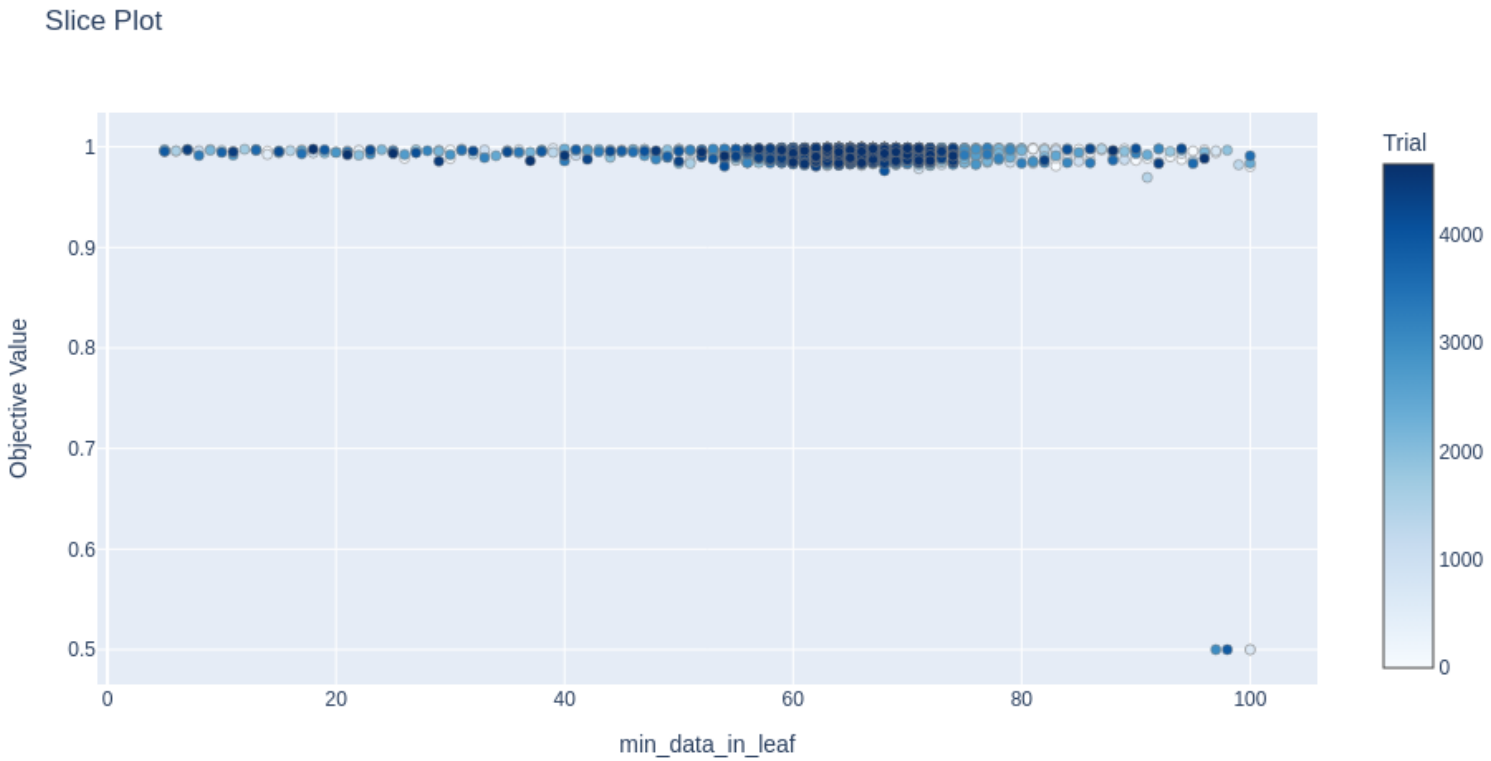


Breast Cancer Wisconsin: LightGBM



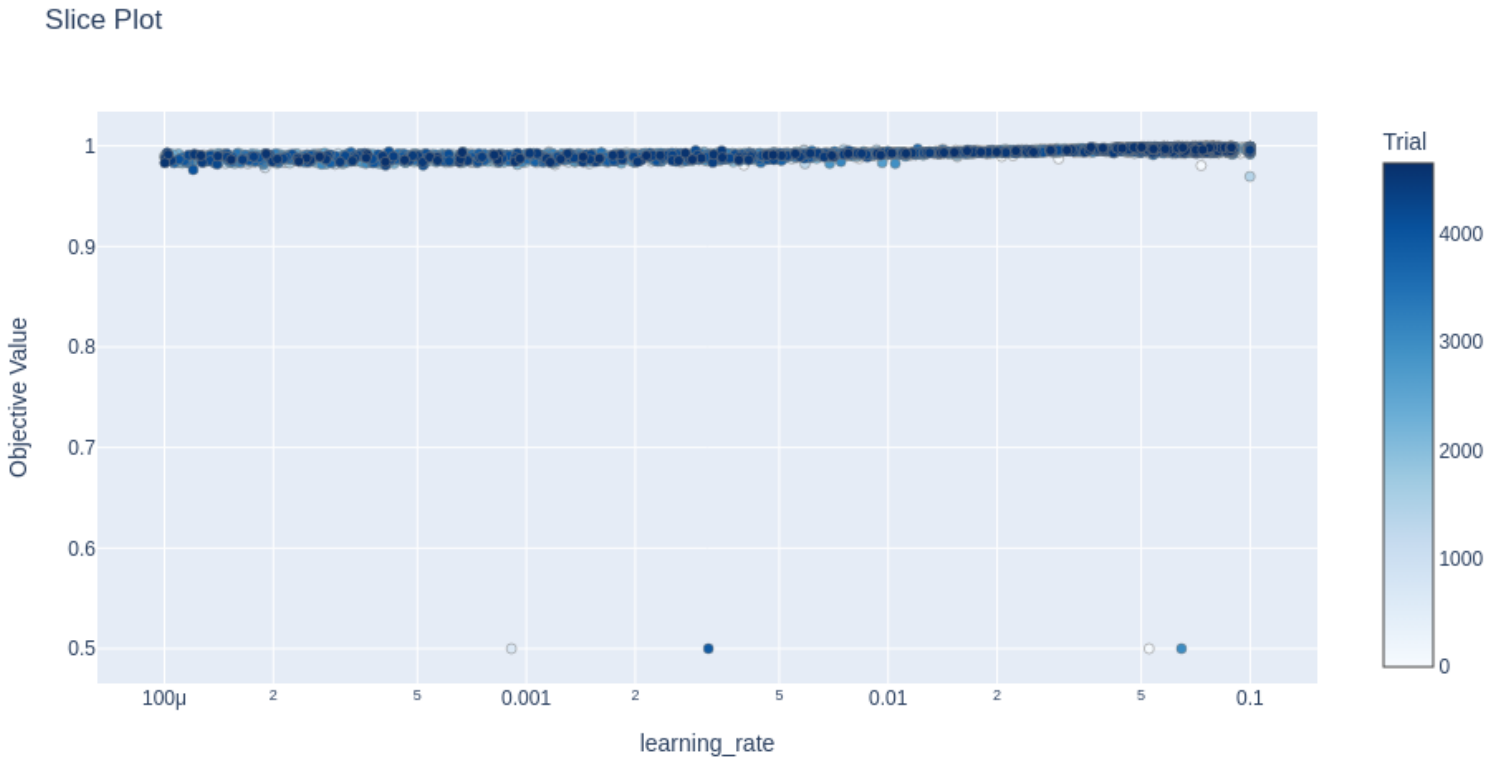
Breast Cancer Wisconsin: LightGBM

'min_data_in_leaf': 66

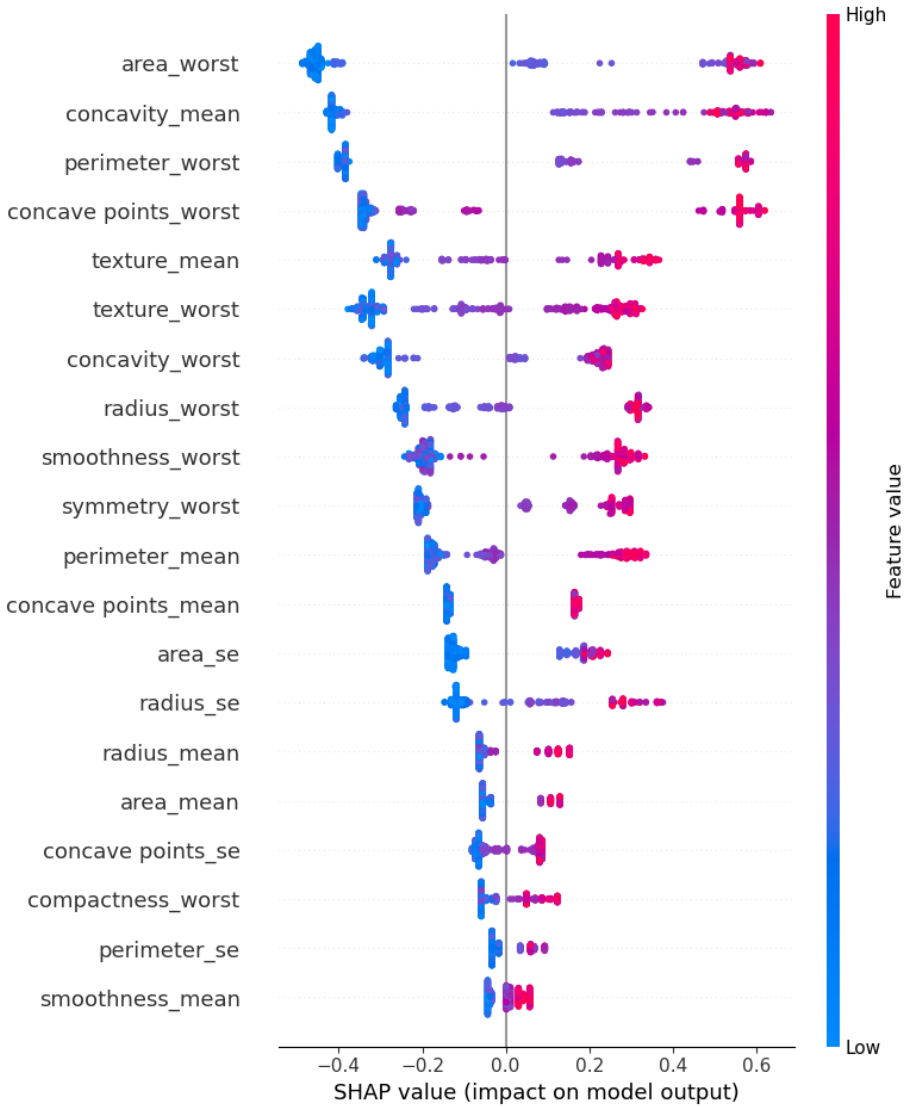


Breast Cancer Wisconsin: LightGBM

'learning_rate': 0.07615521372640538



Breast Cancer Wisconsin: SHAP LightGBM



References

Trevor Hastie and Robert Tibshirani and Jerome Friedman (2009)

The Elements of Statistical Learning

Stuart J. Russell and Peter Norvig (2021)

Artificial Intelligence: A Modern Approach, Global Edition

Max Kuhn and Kjell Johnson (2013)

Applied Predictive Modeling