Time to transplantation

An AI approach to predicting time to transplantation in HT queues

The goal

- What?
  - Predict whether the patient will be in queue longer or less than median time
- Why?
  - Reduce patient stress
  - Planning capabilities
  - Efficient resource allocation
  - Treatment plans
  - Basis for future work
- How?
  - Evaluate many binary classification architectures

Data

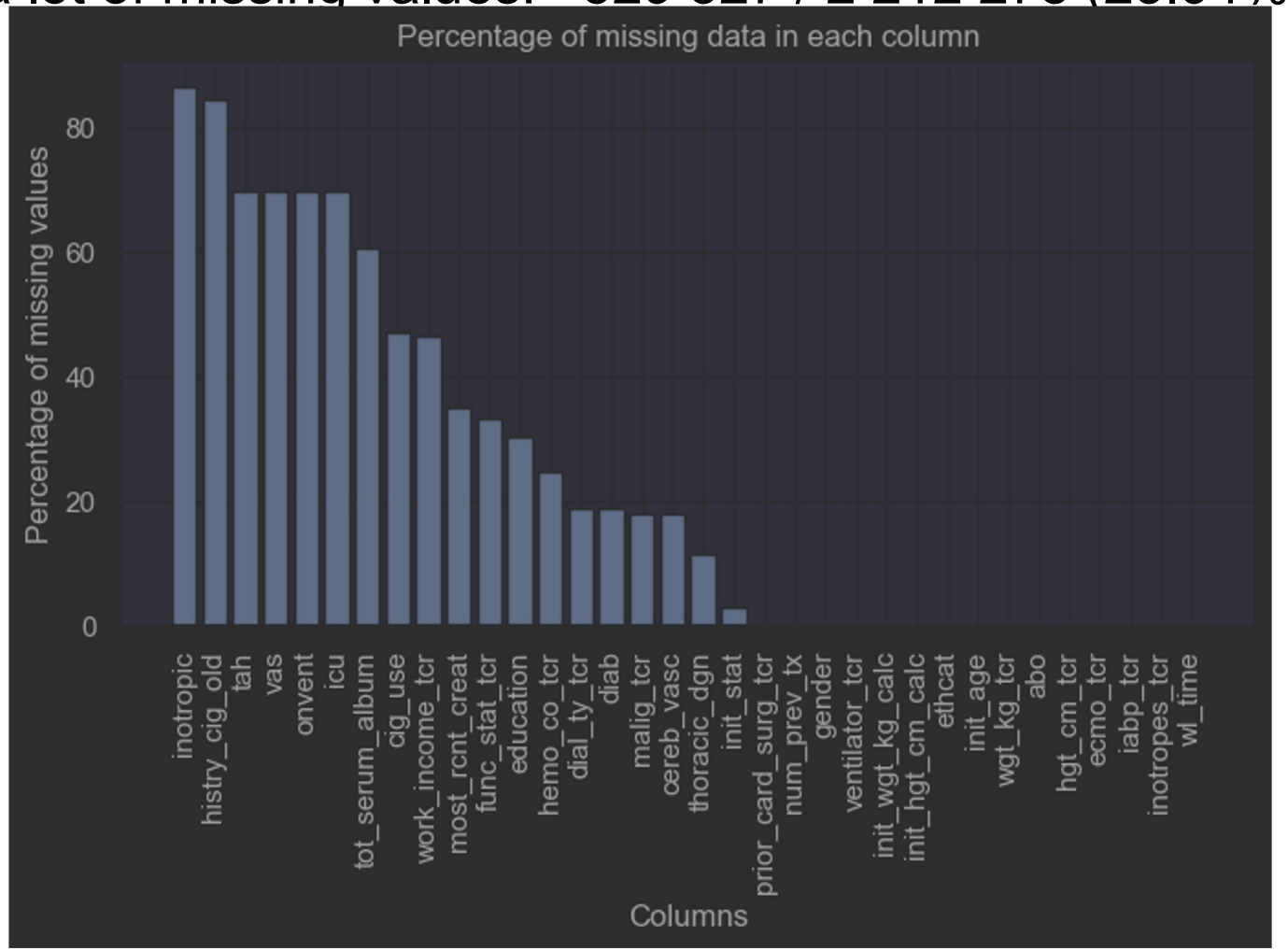
- UNOS database (collect and report on every U.S organ donor, transplant candidate, recipient and outcome)
- 320 variables (columns)
- 120 264 patients in HT queue whom ≈77 410 transplanted

Pre-processing

- Variable (column) selection
  - Only variables measurable when entering queue (320 → 38)
  - Remove empty columns or columns with no variance
  - 25 are categorical, 9 are numerical
- Filtering
  - Remove patients under 18 (120264 → 104 059)
  - Outliers: weight, height, creatinine, albumin, CO (-1772)
  - Only transplanted patients (103 992 patients → 66 839)
- Encoding
  - Blood groups, ethnicities, statuses etc.
- Normalize/scale
- = Median waitlist time of 88 days

Imputation

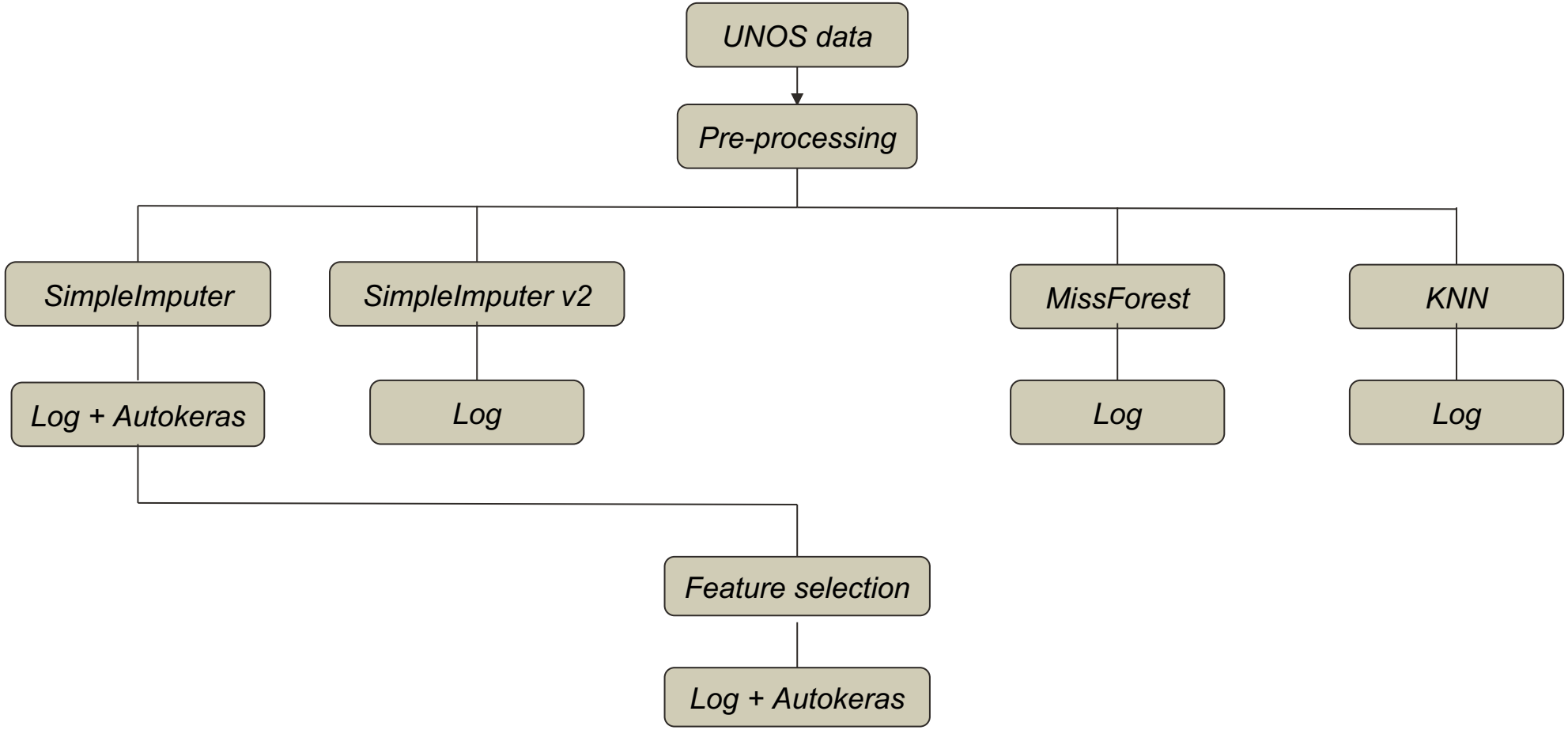
- Assigning values to missing values
- There are a lot of missing values: ≈529 527 / 2 212 278 (23.94%)



Imputation techniques

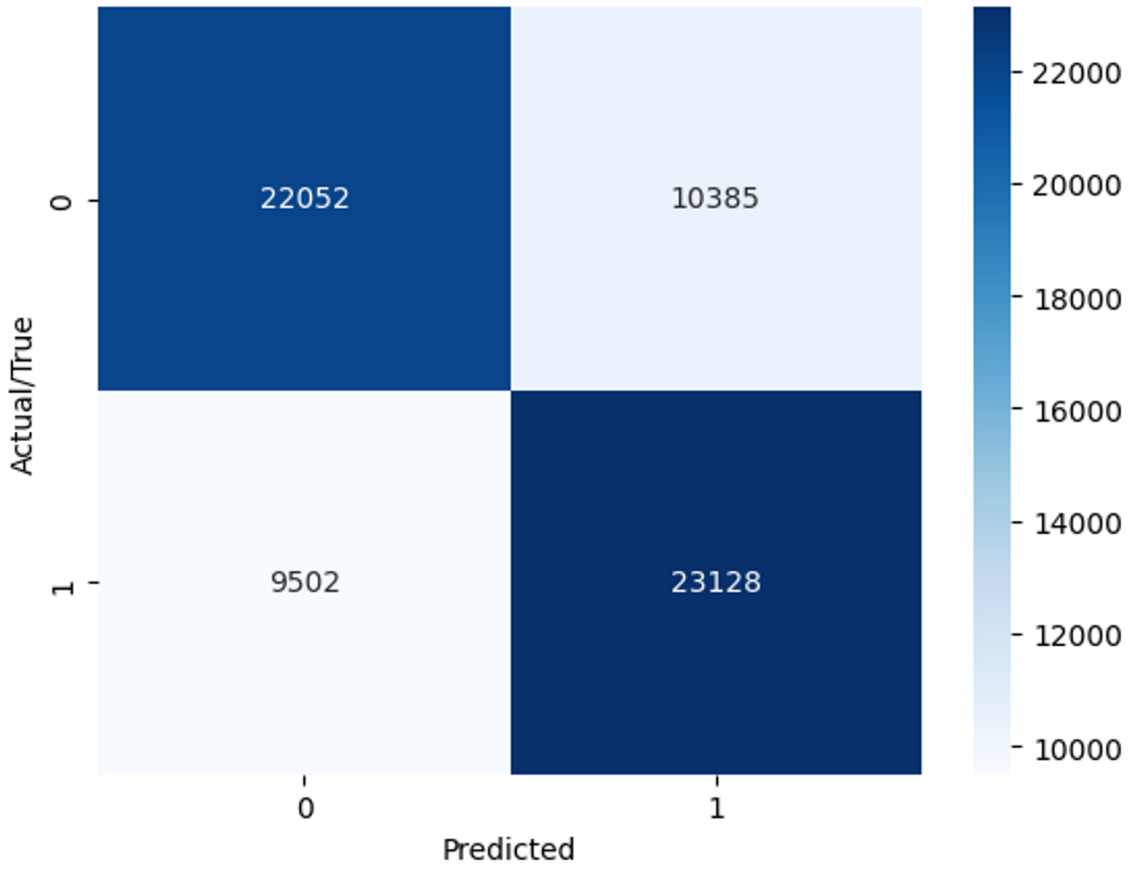
- **SimpleImputer:** “missing” in categorical columns, mean in numerical
- **SimpleImputer v2:** “missing” in categorical, missing indicator for numerical
- **MissForest:** imputes both categorical and numerical, no indicator
- **KNN:** imputes both categorical and numerical, no indicator

Pipelines


UNOS data
Pre-processing
SimpleImputer
SimpleImputer v2
MissForest
KNN
Log + Autokeras
Log
Log
Log
Feature selection
Log + Autokeras

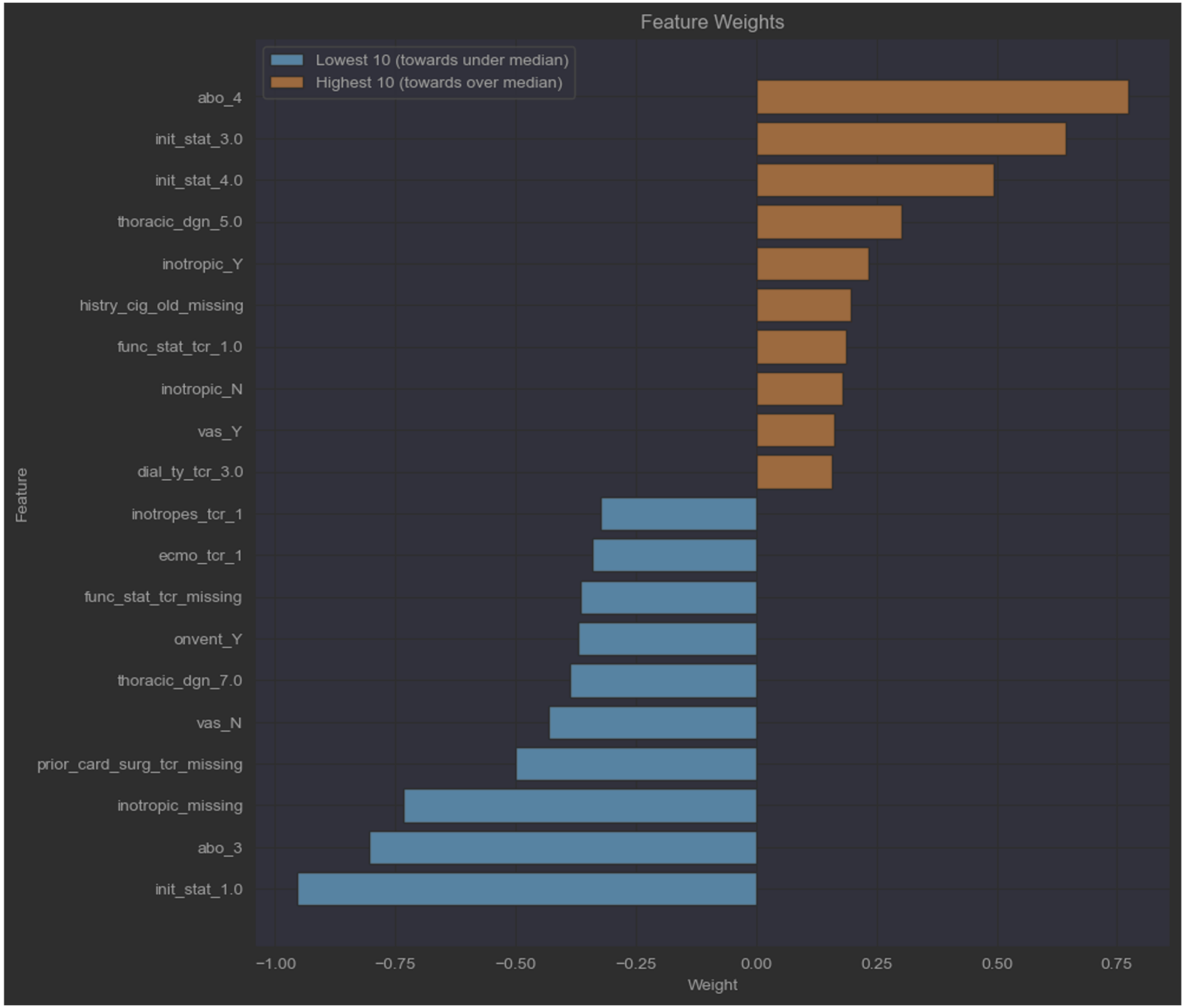
Results pre-feature selection

	Logistic Regression F1 Score	AUROC	Autokeras F1 Score
Simpleimputer*	0.69805	0.76457	0.713234
Simpleimputer_v2	0.68921	0.76463	
missForest	0.69418	0.7569	
KNN	0.69538	0.75671	



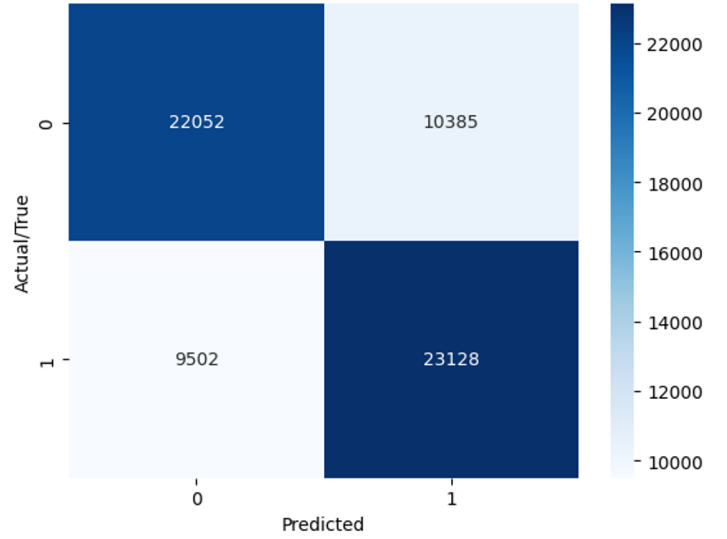
Feature selection

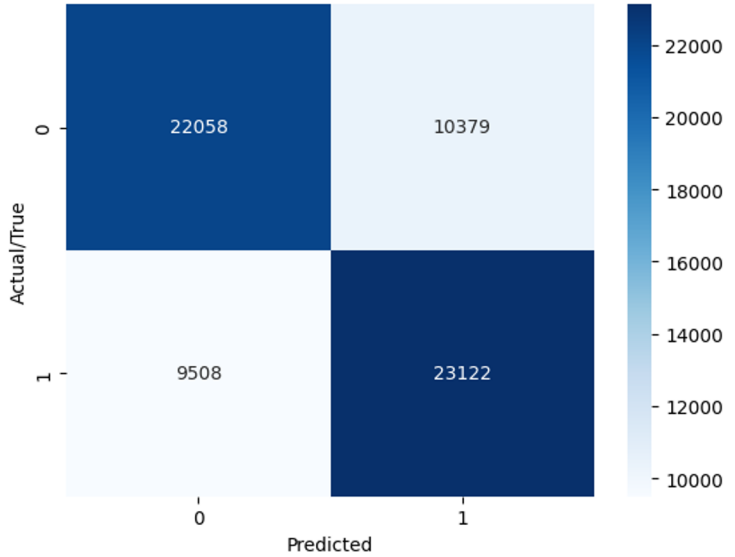
- 99 Features → 85 Features
- Recursive feature selection



Comparison pre/post feature selection

	Logistic Regression F1 Score	AUROC	Autokeras F1 Score
(pre) SimpleImputer	0.69805	0.76457	0.71323
(post) simpleImputer	0.69801	0.76454	0.69933





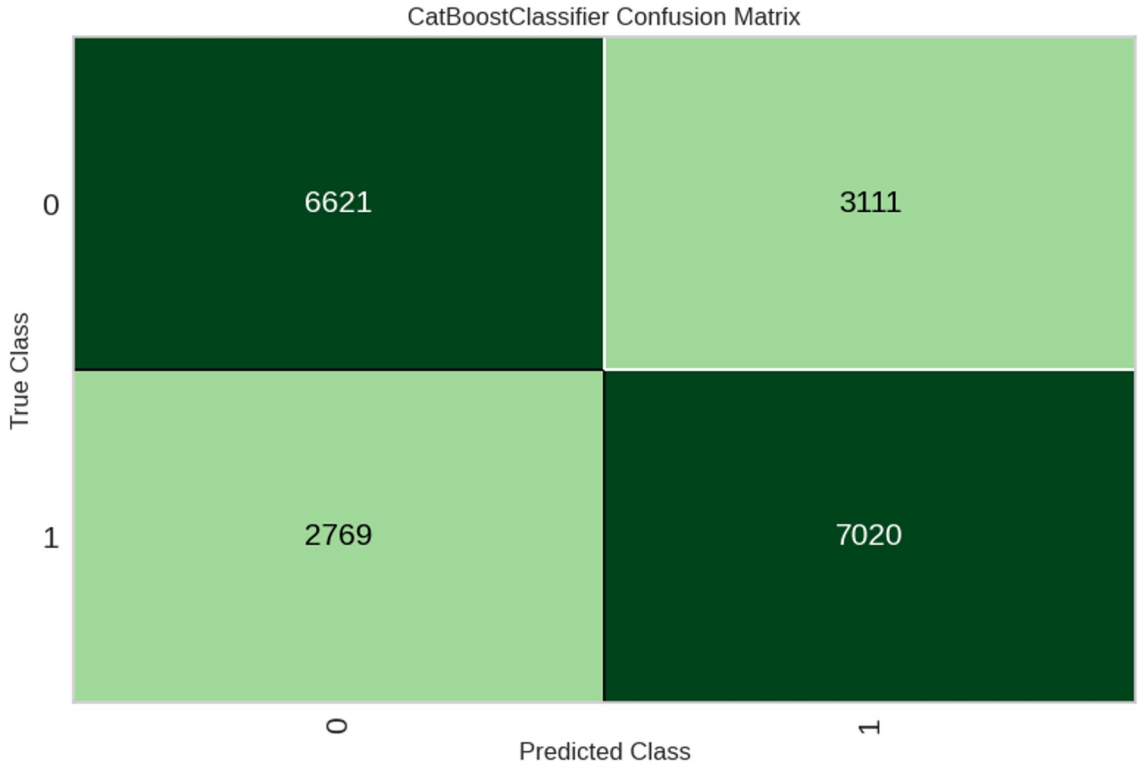
	Description	Value
0	Session id	2058
1	Target	wl_time
2	Target type	Binary
3	Original data shape	(65067, 34)
4	Transformed data shape	(65067, 44)
5	Transformed train set shape	(45546, 44)
6	Transformed test set shape	(19521, 44)
7	Ordinal features	7
8	Numeric features	21
9	Categorical features	12
10	Rows with missing values	100.0%
11	Preprocess	True
12	Imputation type	iterative
13	Iterative imputation iterations	5
14	Numeric iterative imputer	lightgbm
15	Categorical iterative imputer	lightgbm
16	Maximum one-hot encoding	25
17	Encoding method	None
18	Normalize	True
19	Normalize method	zscore
20	Fold Generator	StratifiedKFold
21	Fold Number	10
22	CPU Jobs	-1
23	Use GPU	False
24	Log Experiment	False
25	Experiment Name	clf-default-name
26	USI	f62f

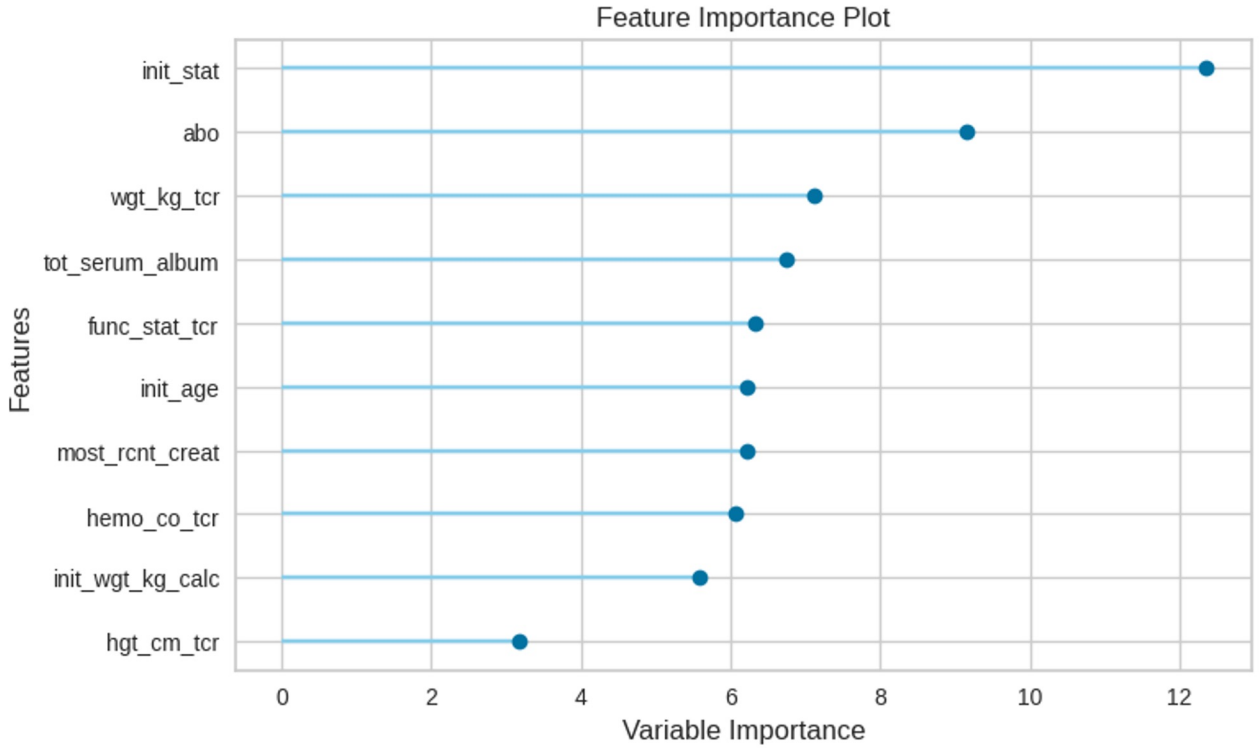
pyCaret - AutoML

- open source, low code, automated

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.7003	0.7653	0.7167	0.6952	0.7058	0.4005	0.4008	10.9920
gbc	Gradient Boosting Classifier	0.6994	0.7650	0.7181	0.6935	0.7056	0.3988	0.3991	2.7850
lightgbm	Light Gradient Boosting Machine	0.6982	0.7643	0.7118	0.6942	0.7029	0.3964	0.3966	1.2920
ada	Ada Boost Classifier	0.6955	0.7590	0.7135	0.6899	0.7015	0.3909	0.3912	1.5040
xgboost	Extreme Gradient Boosting	0.6909	0.7546	0.7047	0.6871	0.6958	0.3818	0.3820	2.4640
rf	Random Forest Classifier	0.6894	0.7503	0.6924	0.6895	0.6909	0.3788	0.3788	2.1550
lr	Logistic Regression	0.6839	0.7436	0.7048	0.6778	0.6910	0.3677	0.3680	13.5260
ridge	Ridge Classifier	0.6838	0.0000	0.7087	0.6763	0.6921	0.3674	0.3679	0.9350
lda	Linear Discriminant Analysis	0.6838	0.7435	0.7087	0.6763	0.6921	0.3675	0.3679	1.1430
et	Extra Trees Classifier	0.6796	0.7400	0.6794	0.6810	0.6802	0.3592	0.3592	2.2520
svm	SVM - Linear Kernel	0.6735	0.0000	0.7044	0.6648	0.6838	0.3469	0.3478	1.0230
knn	K Neighbors Classifier	0.6311	0.6721	0.6482	0.6281	0.6380	0.2621	0.2623	1.6470
nb	Naive Bayes	0.6303	0.6905	0.7809	0.6013	0.6793	0.2599	0.2726	1.1890
dt	Decision Tree Classifier	0.6024	0.6025	0.6036	0.6036	0.6035	0.2048	0.2048	1.1970
qda	Quadratic Discriminant Analysis	0.5914	0.6289	0.7118	0.5769	0.6343	0.1823	0.1904	1.1190
dummy	Dummy Classifier	0.5015	0.5000	1.0000	0.5015	0.6680	0.0000	0.0000	1.1300

pyCaret - AutoML





Conclusion

- We reached a F1 score of 70%
- We found important features for the prediction
- All classification models performed roughly the same
- The imputation method was of less importance as it resulted in same F1 score using logistic regression

Future work

- Replicate study
  - Validate results
  - Different feat selection algorithm
- Multi-classification study
- Predict days

LUND
UNIVERSITY