

Model C: Data-driven AI selection Screening for Clean/Transfused

Table: Sort of P001 Sunderby Hematology Data Update_260326_CM Response: Clean/Transfused Validation: Validation 25_75 by ID
SORTED by Code and Time

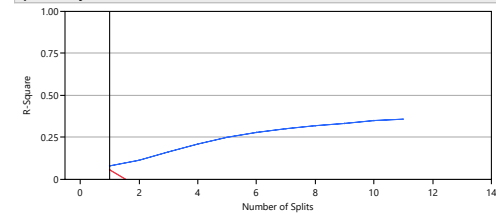
Details

Partition for Clean/Transfused

Validation: Validation 25_75 by ID

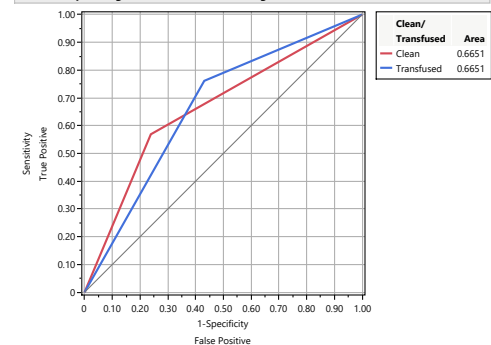
	RSquare	N	Number of Splits
Training	0.080	255	1
Validation	0.057	85	

Split History

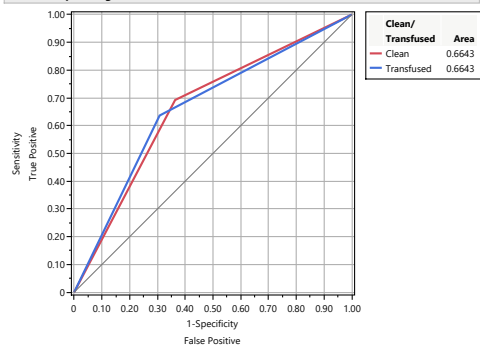


Validation Data in Red

Receiver Operating Characteristic on Training Data



Receiver Operating Characteristic on Validation Data



Bootstrap Forest for Clean/Transfused

Specifications

Target	Clean/Transfused	Training Rows:	255
Validation Column:	Validation 25_75 by ID	Validation Rows:	85
		Test Rows:	0
Number of Trees in the Forest:	100	Number of Terms:	3
Number of Terms Sampled per Split:	2	Bootstrap Samples:	255
		Minimum Splits per Tree:	10
		Minimum Size Split:	5

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.3838	-0.064	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(\emptyset)$
Generalized RSquare	0.5386	-0.120	$(1 - L(0) / L(\text{model}))^{(2/n)} / (1 - L(0) / L(\emptyset))^{(2/n)}$
Mean -Log p	0.3971	0.7105	$\sum y(i) - p(i) / n$
RASE	0.3467	0.5003	$\sqrt{\sum (y(i) - p(i))^2 / n}$
Mean Abs Dev	0.2955	0.4176	$\sum y(i) - p(i) / n$
Misclassification Rate	0.1373	0.4118	$\sum (p(i) \neq \text{Max}(y)) / n$
N	255	85	n

Number of Trees

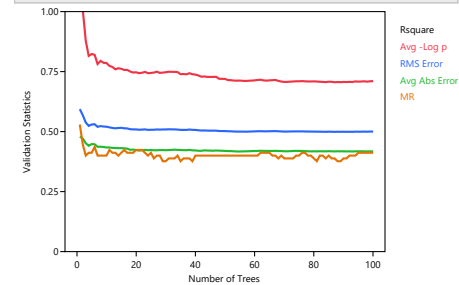
100

Confusion Matrix

Training				Validation			
Actual Clean/Transfused	Predicted Count			Actual Clean/Transfused	Predicted Count		
	Clean	Transfused			Clean	Transfused	
Clean	162	5		Clean	44	8	
Transfused	30	58		Transfused	27	6	

Actual Clean/Transfused	Predicted Rate			Actual Clean/Transfused	Predicted Rate		
	Clean	Transfused			Clean	Transfused	
Clean	0.970	0.030		Clean	0.846	0.154	
Transfused	0.341	0.659		Transfused	0.818	0.182	

Cumulative Validation

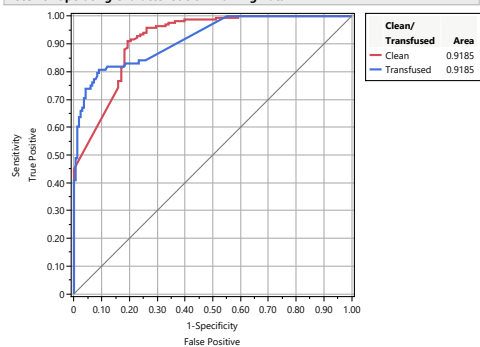


Model C: Data-driven AI selection Screening for Clean/Transfused

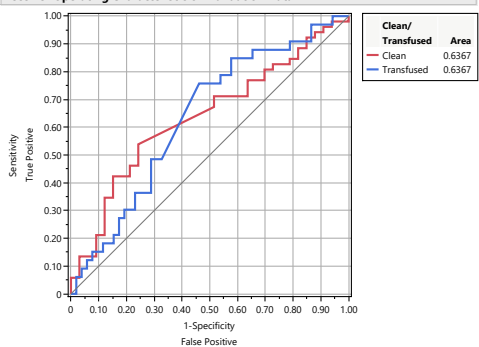
Details

Bootstrap Forest for Clean/Transfused

Receiver Operating Characteristic on Training Data



Receiver Operating Characteristic on Validation Data



Boosted Tree for Clean/Transfused

Specifications

Target	Clean/Transfused	Number of training rows:	255
Validation Column:	Validation 25, 75 by ID	Number of validation rows:	85
Number of Layers:	27		
Splits per Tree:	2		
Learning Rate:	0.092		
Overfit Penalty:	0.0001		

Overall Statistics

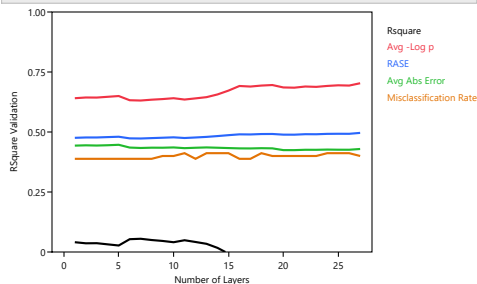
Measure	Training	Validation	Definition
Entropy RSquare	0.2092	-0.053	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.3262	-0.099	$(1 - L(0) / L(\text{model}))^n / (2/n) / (1 - L(0) / L(0))^n / (2/n)$
Mean -Log p	0.5096	0.7032	$\sum -\log(p[j]) / n$
RASE	0.4090	0.4963	$\sqrt{\sum (y[j] - p[j])^2 / n}$
Mean Abs Dev	0.3702	0.4292	$\sum y[j] - p[j] / n$
Misclassification Rate	0.2275	0.4000	$\sum (p[j] \neq \text{Max}) / n$
N	255	85	n

Confusion Matrix

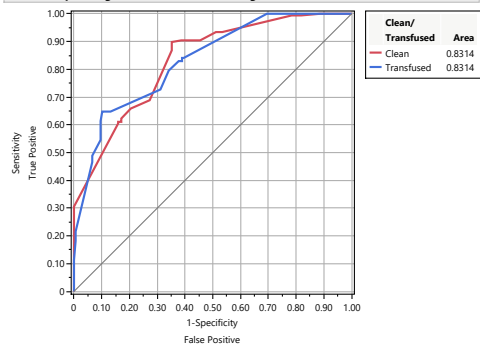
Training			Validation		
Actual	Predicted Count		Actual	Predicted Count	
Clean/Transfused	Clean	Transfused	Clean/Transfused	Clean	Transfused
Clean	156	11	Clean	46	6
Transfused	47	41	Transfused	28	5

Actual	Predicted Rate		Actual	Predicted Rate	
Clean/Transfused	Clean	Transfused	Clean/Transfused	Clean	Transfused
Clean	0.934	0.066	Clean	0.885	0.115
Transfused	0.534	0.466	Transfused	0.848	0.152

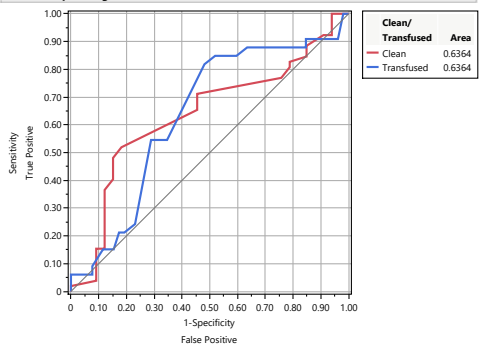
Cumulative Validation



Receiver Operating Characteristic on Training Data



Receiver Operating Characteristic on Validation Data



Model C: Data-driven AI selection Screening for Clean/Transfused

Details

Neural

Validation Column: Validation 25, 75 by ID

Model Launch

Hidden Layer Structure

Number of nodes of each activation type

Activation	Sigmoid	Identity	Radial
Layer	TanH	Linear	Gaussian
First	3	0	0
Second	0	0	0

Second layer is closer to X's in two layer models.

Boosting

Fit an additive sequence of models scaled by the learning rate.

Number of Models	0
Learning Rate	0.1

Fitting Options

☐ Transform Covariates

Penalty Method

Number of Tours

Model NTanH(3)NBoost(1)

Training

Clean/Transfused

Measures	Value
Generalized RSquare	3.314e-15
Entropy RSquare	1.776e-15
RASE	0.4873629
Mean Abs Dev	0.4750453
Misclassification Rate	0.3882979
-LogLikelihood	125.58036
Sum Freq	188

Confusion Matrix

Actual	Predicted Count	
Clean/Transfused	Clean	Transfused
Clean	115	0
Transfused	73	0

Confusion Rates

Actual	Predicted Rate	
Clean/Transfused	Clean	Transfused
Clean	1.000	0.000
Transfused	1.000	0.000

Validation

Clean/Transfused

Measures	Value
Generalized RSquare	-0.014821
Entropy RSquare	-0.006315
RASE	0.4758035
Mean Abs Dev	0.4639116
Misclassification Rate	0.3384615
-LogLikelihood	41.946538
Sum Freq	65

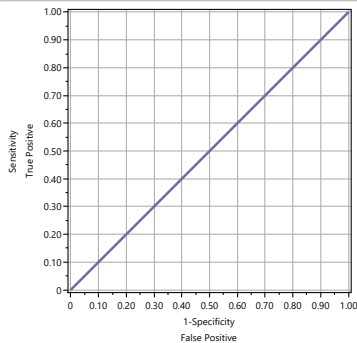
Confusion Matrix

Actual	Predicted Count	
Clean/Transfused	Clean	Transfused
Clean	43	0
Transfused	22	0

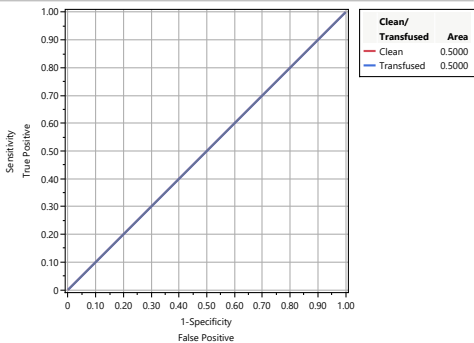
Confusion Rates

Actual	Predicted Rate	
Clean/Transfused	Clean	Transfused
Clean	1.000	0.000
Transfused	1.000	0.000

Receiver Operating Characteristic on Training Data



Receiver Operating Characteristic on Validation Data



Model C: Data-driven AI selection Screening for Clean/Transfused

Details

Support Vector Machine												
Model Comparison												
Show	Method	Kernel Function	Cost	Gamma	# SV	Training Misclassification Rate	Validation Misclassification Rate	Validation Generalized RSquare	Probability Threshold	Threshold Validation Misclassification Rate	Best	
☒	Model 1	Radial Basis Function	1	0.33333	145	0.26064	0.26154	0.04033	0.01	0.33846	Smallest Misclassification Rate Largest RSquare	

Validation Column: Validation 25_75 by ID

Model Launch

Kernel Function

Radial Basis Function

Tuning Design

Cost1

Gamma0.333333

Support Vector Machine Model 1

Model Summary		Estimation Details	
Response	Clean/Transfused	Cost	1
Validation Method	Validation Column	Gamma	0.33333
Kernel Function	Radial Basis Function		
Measure	Training	Validation	
Number of rows	188	65	
Sum of Frequencies	188	65	
Misclassification Rate	0.2606383	0.2615385	
Number of Support Vectors	145	145	

Fit Details		
Measure	Training	Validation Definition
Entropy RSquare	0.2234	0.0231 1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3501	0.0403 (1-L(0)/L(model))^(2/n)/(1-L(0)^(2/n))
Mean -Log p	0.5187	0.6252 $\sum -\log(p_{ij})/n$
RASE	0.4140	0.4575 $\sqrt{\sum (y_{ij}-\hat{p}_{ij})^2/n}$
Mean Abs Dev	0.3447	0.3819 $\sum y_{ij}-\hat{p}_{ij} /n$
Misclassification Rate	0.2606	0.2462 $\sum (p_{ij}-p_{Max})/n$
N	188	65 n

Confusion Matrix

Set Probability Threshold

Probability Threshold

0.01

Training

Actual	Predicted Rate		Misclassification Rate
Clean/Transfused	Clean	Transfused	
Clean	1.000	0.000	0.3883
Transfused	1.000	0.000	

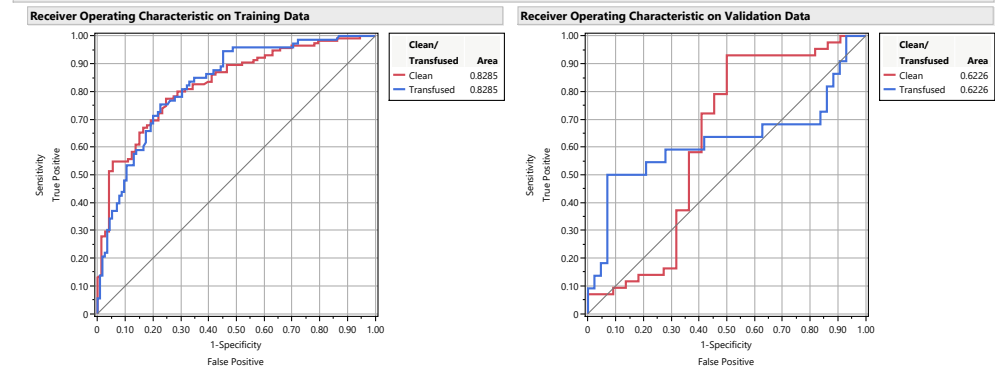
Actual	Predicted Count	
Clean/Transfused	Clean	Transfused
Clean	115	0
Transfused	73	0

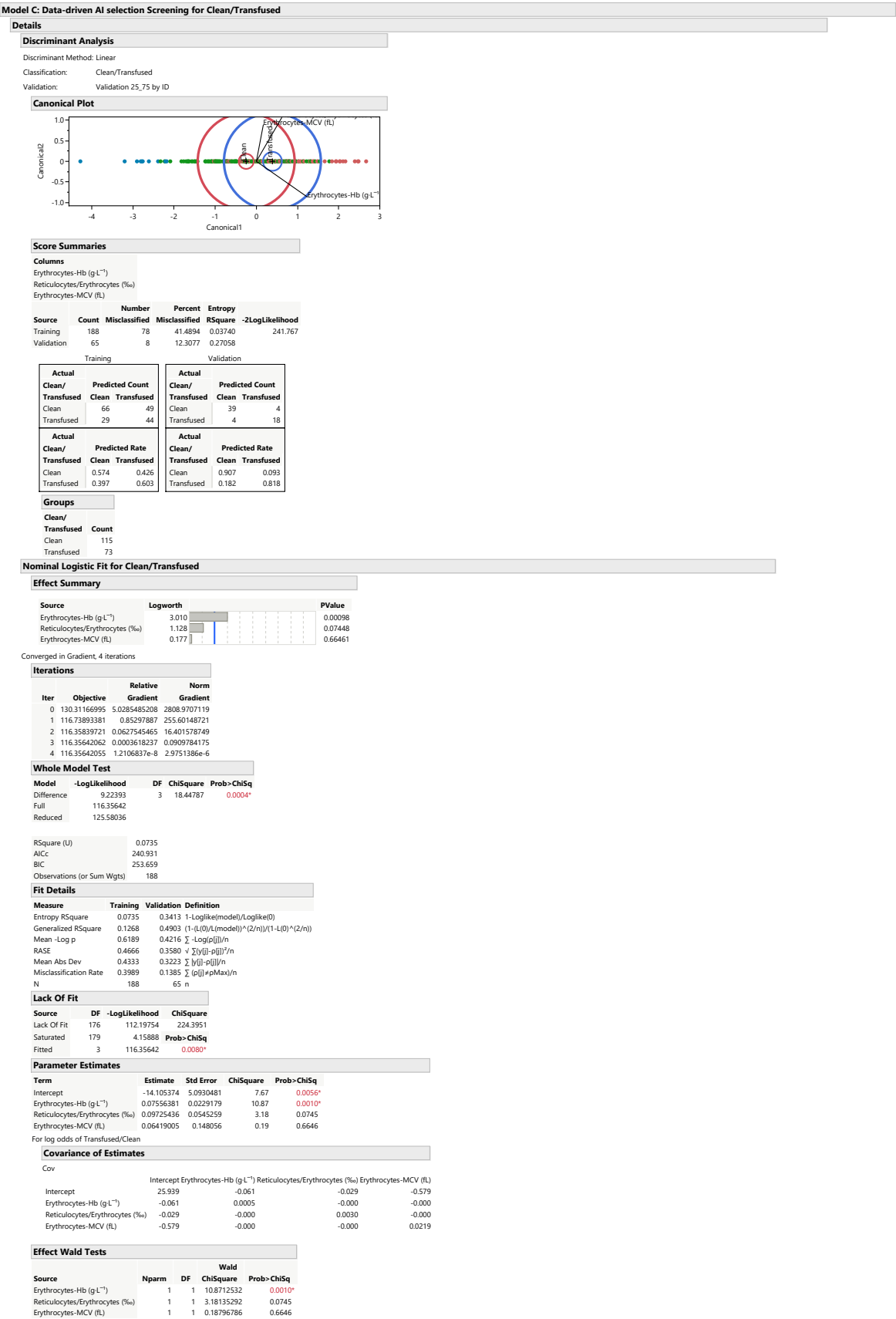
Validation

Actual	Predicted Rate		Misclassification Rate
Clean/Transfused	Clean	Transfused	
Clean	1.000	0.000	0.3385
Transfused	1.000	0.000	

Actual	Predicted Count	
Clean/Transfused	Clean	Transfused
Clean	43	0
Transfused	22	0

ROC Curve for Clean/Transfused = Clean



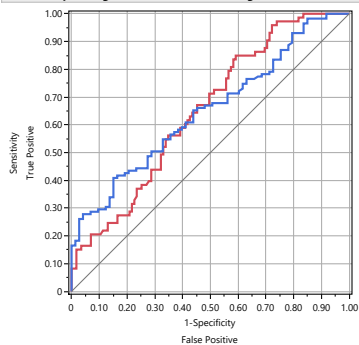


Model C: Data-driven AI selection Screening for Clean/Transfused

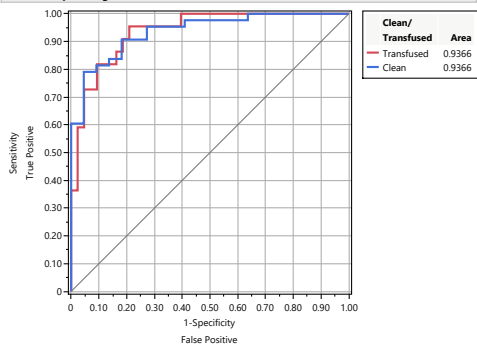
Details

Nominal Logistic Fit for Clean/Transfused

Receiver Operating Characteristic on Training Data



Receiver Operating Characteristic on Validation Data



Generalized Regression for Clean/Transfused = Transfused

Model Comparison

Show	Response Distribution	Estimation Method	Validation Method	Nonzero Parameters	AICc	BIC	Generalized RSquare	Validation Generalized RSquare
☒	Binomial	Lasso	Validation Column	4	240.93142	253.65861	0.1268035	0.4902401

Model Launch

Response Distribution

Estimation Method

☐ Adaptive

Validation Method

☐ Early Stopping

Validation Column: Validation 25_75 by ID

Binomial Lasso with Validation Column

Model Summary

Response: Clean/Transfused
Distribution: Binomial
Estimation Method: Lasso
Validation Method: Validation Column
Probability Model Link: Logit
Validation Column: Validation 25_75 by ID

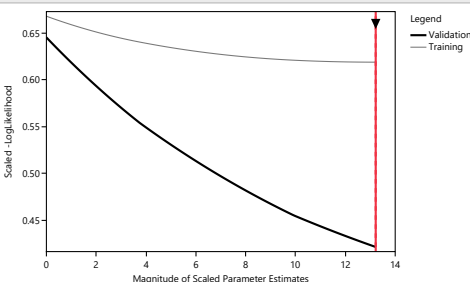
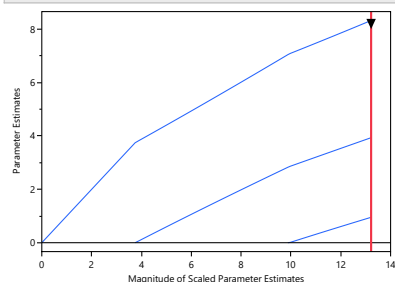
Estimation Details

Number of Grid Points: 150
Minimum Penalty Fraction: 1e-4
Grid Scale: Square Root

Measure

	Training	Validation
Number of rows	188	65
Sum of Frequencies	188	65
-LogLikelihood	116.35642	27.402704
Number of Parameters	4	4
BIC	253.65861	71.502957
AICc	240.93142	63.472075
Generalized RSquare	0.1268035	0.4902401
Lambda Penalty	0.0001821	

Solution Path



Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-14.10237	4.4445832	10.067516	0.0015*	-22.8136	-5.39115
Erythrocytes-Hb (g L ⁻¹)	0.0755546	0.0203707	13.756531	0.0002*	0.0356287	0.1154805
Reticulocytes/Erythrocytes (%)	0.0972332	0.0513298	3.5883097	0.0582	-0.003371	0.1978379
Erythrocytes-MCV (fL)	0.0641406	0.1430939	0.2009199	0.6540	-0.216318	0.3445995

Training

Method	N	Sum Wgt	Entropy RSquare	Misclassification Rate	AUC	RASE	Generalized RSquare
Bootstrap Forest	255	255.000	0.3838	0.1373	0.9185	0.34673	0.5386
Support Vector Machines	188	188.000	0.2234	0.2606	0.8285	0.41399	0.3501
Boosted Tree	255	255.000	0.2092	0.2275	0.8314	0.40902	0.3262
Decision Tree	255	255.000	0.0803	0.3451	0.6651	0.45116	0.1357
Nominal Logistic	188	188.000	0.0735	0.3989	0.6508	0.46659	0.1268
Generalized Regression Lasso	188	188.000	0.0735	0.3989	0.6508	0.46659	0.1268
Discriminant	188	188.000	0.0374	0.4149			
Neural Boosted	188	188.000	0.0000	0.3883	0.5000	0.48736	0.0000

Validation

Method	N	Sum Wgt	Entropy RSquare	Misclassification Rate	AUC	RASE	Generalized RSquare
Nominal Logistic	65	65.0000	0.3413	0.1385	0.9366	0.35801	0.4903
Generalized Regression Lasso	65	65.0000	0.3413	0.1385	0.9366	0.35802	0.4902
Discriminant	65		0.2706	0.1231			
Decision Tree	85	85.0000	0.0572	0.3882	0.6643	0.46750	0.0997
Support Vector Machines	65	65.0000	0.0231	0.2462	0.6226	0.45754	0.0403
Neural Boosted	65	65.0000	-0.008	0.3385	0.5000	0.47580	-0.015
Boosted Tree	85	85.0000	-0.053	0.4000	0.6364	0.49627	-0.099
Bootstrap Forest	85	85.0000	-0.064	0.4118	0.6367	0.50029	-0.120

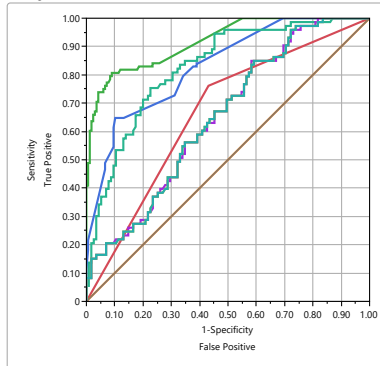
+ indicates that the method uses validation data to tune model.

Model C: Data-driven AI selection Screening for Clean/Transfused

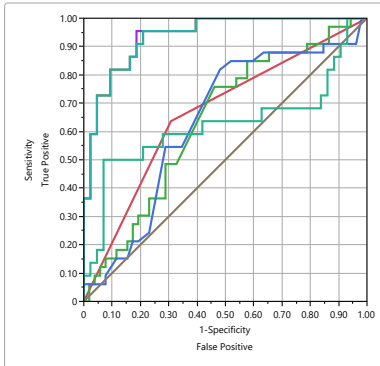
Sum Freq is suppressed when it is the same as N.

ROC Curve

Training



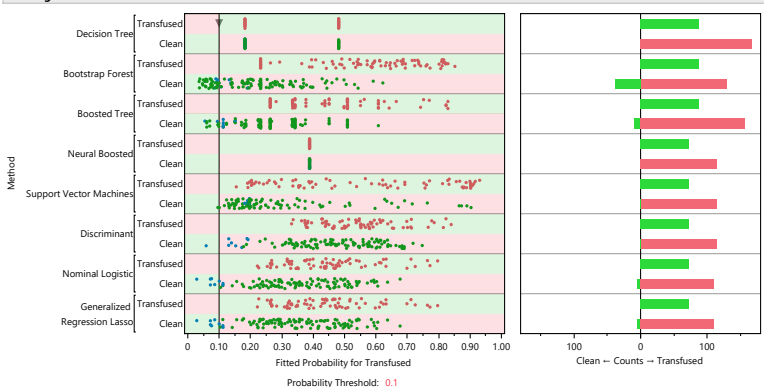
Validation



Method	AUC	
	Training	Validation
Decision Tree	0.6651	0.6643
Bootstrap Forest	0.9185	0.6367
Boosted Tree	0.8314	0.6364
Neural Boosted	0.5000	0.5000
Support Vector Machines	0.8205	0.6236
Discriminant	0.6508	0.9376
Nominal Logistic	0.6508	0.9366
Generalized Regression Lasso	0.6508	0.9366

Decision Thresholds

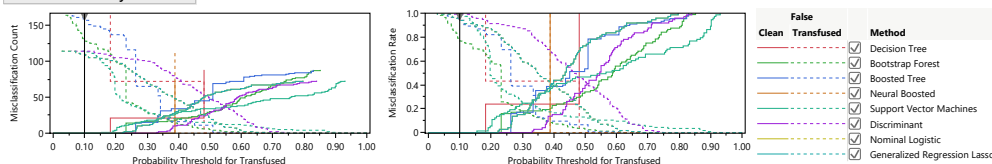
Training



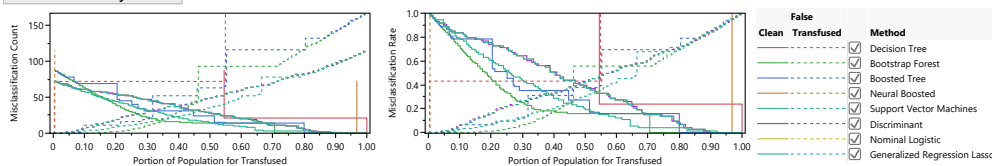
Method	Actual	Predicted		Rates	
		Clean	Transfused	Clean	Transfused
Decision Tree	Transfused	0	88	0.000	1.000
	Clean	0	167	0.000	1.000
Bootstrap Forest	Transfused	0	88	0.000	1.000
	Clean	38	129	0.228	0.772
Boosted Tree	Transfused	0	88	0.000	1.000
	Clean	10	157	0.060	0.940
Neural Boosted	Transfused	0	73	0.000	1.000
	Clean	0	115	0.000	1.000
Support Vector Machines	Transfused	0	73	0.000	1.000
	Clean	1	114	0.009	0.991
Discriminant	Transfused	0	73	0.000	1.000
	Clean	1	114	0.009	0.991
Nominal Logistic	Transfused	0	73	0.000	1.000
	Clean	5	110	0.043	0.957
Generalized Regression Lasso	Transfused	0	73	0.000	1.000
	Clean	5	110	0.043	0.957

Graphs

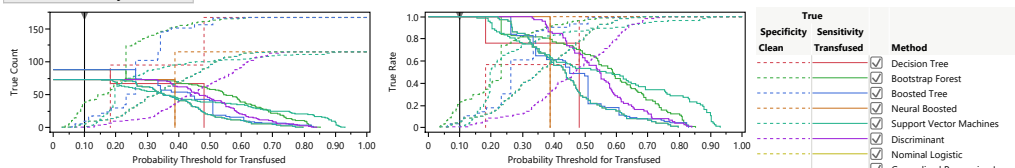
False Classification by Threshold



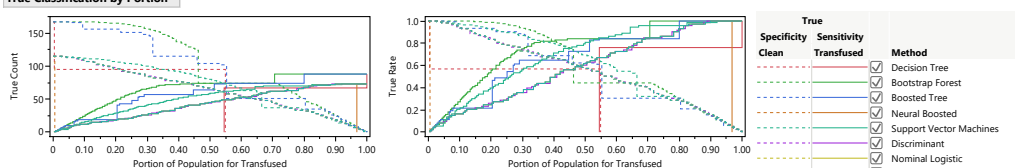
False Classification by Portion



True Classification by Threshold



True Classification by Portion



Model C: Data-driven AI selection Screening for Clean/Transfused

Decision Thresholds

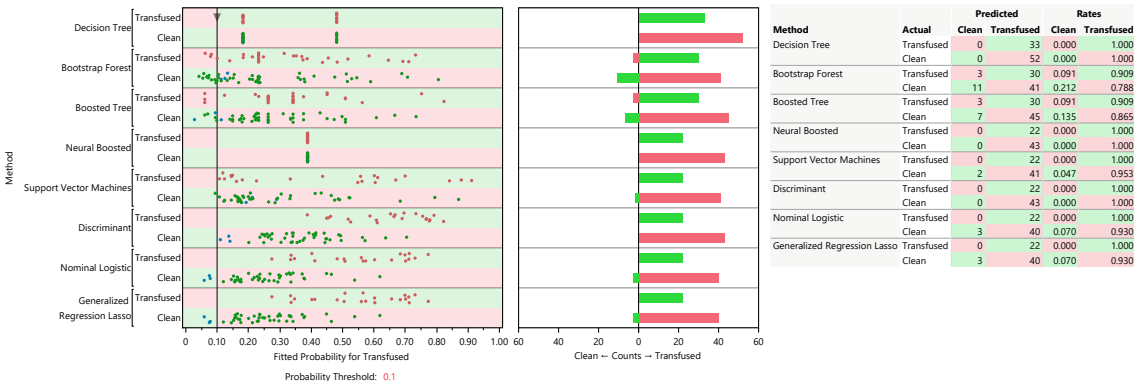
Training

Metrics

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Decision Tree	88	0	167	0	1.0000	0.0000	0.3451	0.3451	0.5131	-
Bootstrap Forest	88	0	129	38	1.0000	0.2275	0.4055	0.4941	0.5777	0.3038
Boosted Tree	88	0	157	10	1.0000	0.0599	0.3592	0.3843	0.5285	0.1467
Neural Boosted	73	0	115	0	1.0000	0.0000	0.3883	0.3883	0.5594	-
Support Vector Machines	73	0	114	1	1.0000	0.0087	0.3904	0.3936	0.5615	0.0583
Discriminant	73	0	114	1	1.0000	0.0087	0.3904	0.3936	0.5615	0.0583
Nominal Logistic	73	0	110	5	1.0000	0.0435	0.3989	0.4149	0.5703	0.1317
Generalized Regression Lasso	73	0	110	5	1.0000	0.0435	0.3989	0.4149	0.5703	0.1317

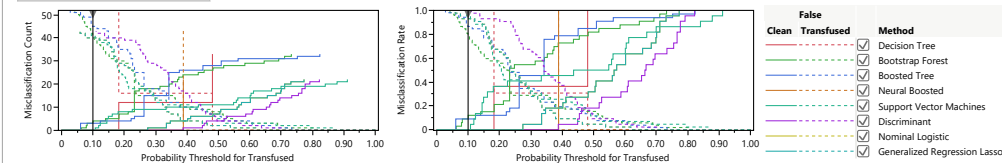
Where Transfused is considered Positive. Probability Threshold: 0.100

Validation

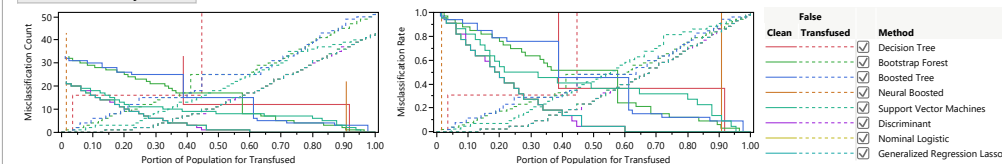


Graphs

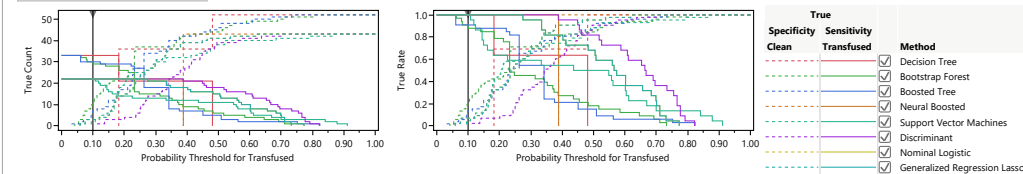
False Classification by Threshold



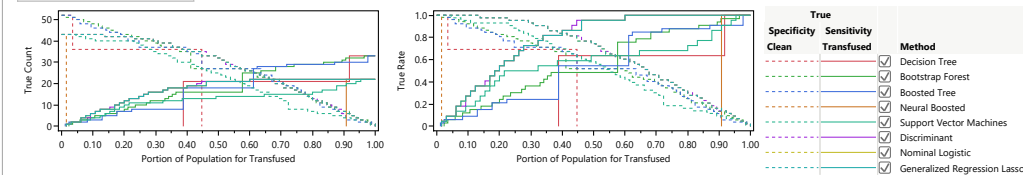
False Classification by Portion



True Classification by Threshold



True Classification by Portion



Metrics

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Decision Tree	33	0	52	0	1.0000	0.0000	0.3882	0.3882	0.5593	-
Bootstrap Forest	30	3	41	11	0.9091	0.2115	0.4225	0.4824	0.5769	0.1585
Boosted Tree	30	3	45	7	0.9091	0.1346	0.4	0.4353	0.5556	0.0661
Neural Boosted	22	0	43	0	1.0000	0.0000	0.3385	0.3385	0.5057	-
Support Vector Machines	22	0	41	2	1.0000	0.0465	0.3492	0.3692	0.5176	0.1274
Discriminant	22	0	43	0	1.0000	0.0000	0.3385	0.3385	0.5057	-
Nominal Logistic	22	0	40	3	1.0000	0.0698	0.3548	0.3846	0.5238	0.1573
Generalized Regression Lasso	22	0	40	3	1.0000	0.0698	0.3548	0.3846	0.5238	0.1573

Where Transfused is considered Positive. Probability Threshold: 0.100