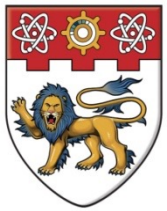


NANYANG TECHNOLOGICAL UNIVERSITY

Unlocking the prognostic potential of blood-based gene expression data from ADNI and ANMERGE

CBI / LKC School of Medicine

Eric LIM JIT KAI
Mohammad Neamul Kabir
Han Wenhao

Feb/01/2023

Table of content

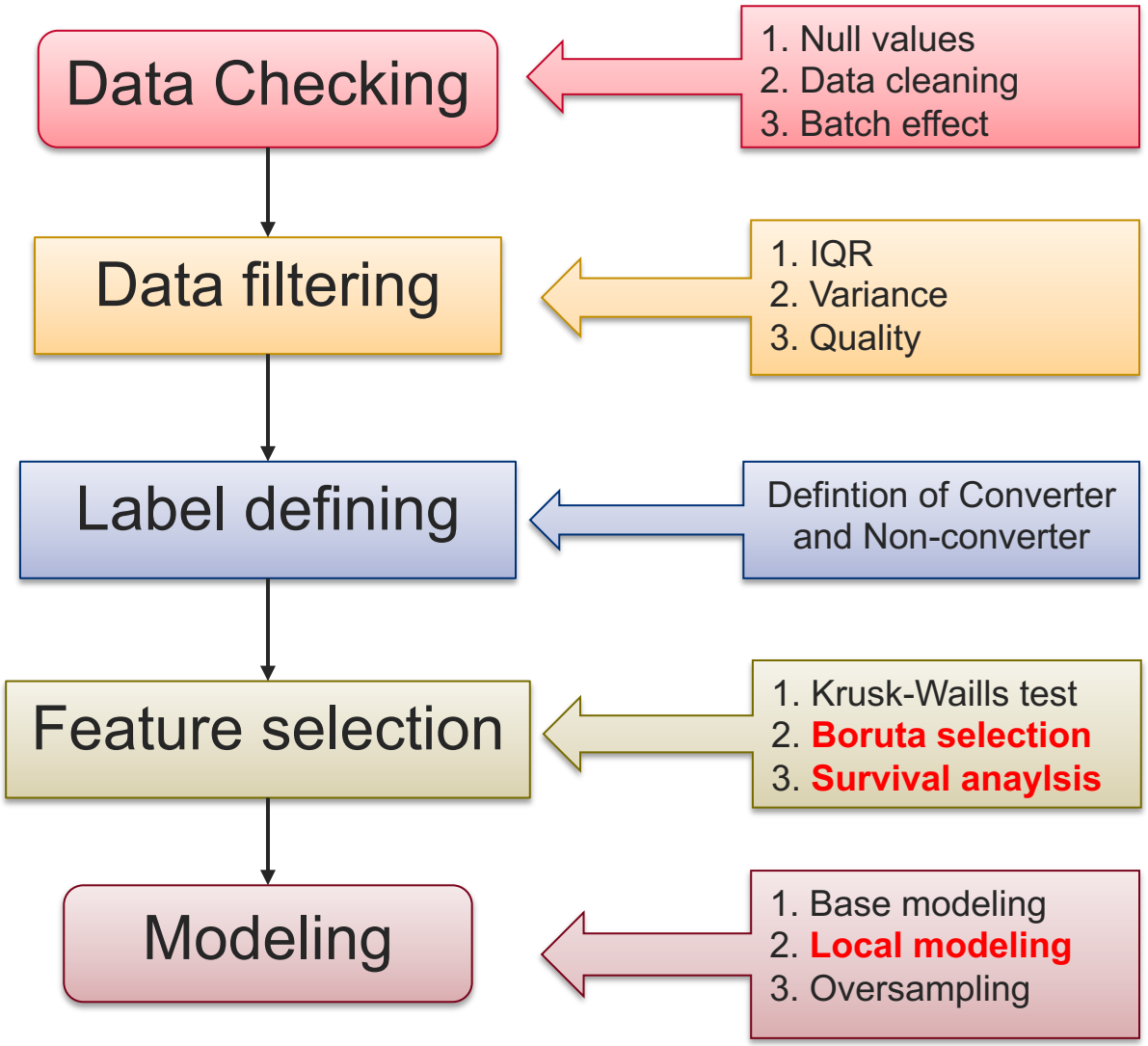
- Story Line of the project
- Methodology
- Modeling and optimization
- Work to be done

Story Line of the project

Main proposals:

1. Presence of batch effect. Metadata allow us to check their technical factors and confounding factors.

2. Derive predictive ability from the conversion duration data and APOE4 alleles.
(via survival analysis, Boruta and local modeling)



NANYANG TECHNOLOGICAL UNIVERSITY

Methodology

- Data cleaning
- Definition of conversion (labeling)
- Survival analysis:

  Cox proportional-hazards model
- Set the threshold for gene selection($p <= 0.05$)
- Split the training set to three based on the value of APOE4
- Boruta gene selection

Definition of Conversion(class label)

Definition of class label				
Definition	**start diagnosis**	**end diagnosis**	**Class label**	**count**
Converter	MCI	Dementia	1	95 (29%)
Non-converter	MCI	MCI	0	234 (71%)

Dataset for local modeling			
class label	**APOE4 = 0**	**APOE4 = 1**	**APOE4 = 2**
1	49	41	5
0	145	72	17
Sum	194	113	22

Apply Boruta for feature selection

- We considered genes with a p-value lower than 0.05 for survival analysis as significant genes and applied them for model training.
- There are 1715 genes are identified as significant genes.

The shape of dataset (num of sample, num of genes)				
	Base modeling	**Local modeling**		
	All samples	**APOE4 = 0**	**APOE4 =1**	**APOE4 = 2**
Before Boruta	(329, 1715)	(194, 1715)	(113, 1715)	(22, 1715)
After Boruta	(329, 135)	(194, 80)	(113, 113)	(22, 50)

Modeling and optimization

- Applying Boruta for gene selection
- Model optimization for RF and SVM
- Simple classifiers testing(LG, KNN, NB)
- Local modeling for all above

Hyperparameter tuning for RF model

Optimal hyperparameter and accuracy				
	dataset	**n_estimators**	**max_depth**	**accuracy**
Base modeling	**All**	50	10	0.73
	All_boruta	80	20	**0.75**
Local modeling	**APOE4_0**	20	10	0.75
	APOE4_1	30	10	0.68
	APOE4_2	10	10	0.77
	APOE4_0_boruta	20	10	**0.78**
	APOE4_1_boruta	40	10	**0.83**
	APOE4_2_boruta	20	10	**0.85**

Hyperparameter tuning for SVM model

Optimal hyperparameter and accuracy (rbf kernel)				
	dataset	**C**	**gamma**	**accuracy**
Base modeling	**All**	100	0.0001	0.74
	All_boruta	1.0	0.01	**0.84**
Local modeling	**APOE4_0**	10	0.0001	0.78
	APOE4_1	100	0.0001	0.73
	APOE4_2	0.01	1e-9	0.77
	APOE4_0_boruta	10	0.01	**0.86**
	APOE4_1_boruta	1.0	0.01	**0.83**
	APOE4_2_boruta	1.0	0.01	**1.00**

Hyperparameter tuning for SVM model

Optimal hyperparameter and accuracy (Linear kernel)				
	dataset	**C**	**accuracy (linear)**	**accuracy (rbf)**
Base modeling	**All**	0.01	0.73	0.74
	All_boruta	0.01	**0.79**	**0.84**
Local modeling	**APOE4_0**	0.01	0.75	0.78
	APOE4_1	0.01	0.70	0.73
	APOE4_2	0.01	0.72	0.77
	APOE4_0_boruta	0.1	**0.83**	**0.86**
	APOE4_1_boruta	0.1	**0.79**	**0.83**
	APOE4_2_boruta	0.1	**1.00**	**1.00**

Boruta improve the base model performance

Comparison of evaluation metrics for the optimal base models										
	accuracy		precision		recall		F1 score		AUC	
RF	0.71	0.76	0.36	0.83	0.09	0.24	0.15	0.36	0.63	0.79
SVM (rbf)	0.74	**0.83**	0.57	**0.89**	0.44	0.49	0.49	0.62	0.74	**0.86**
SVM (linear)	0.73	0.81	0.55	0.70	0.44	0.61	0.49	**0.64**	0.74	**0.86**
LG	0.73	0.74	0.54	0.57	0.42	0.53	0.47	0.54	0.75	0.80
KNN	0.70	0.73	0.50	0.72	0.07	0.16	0.12	0.25	0.55	0.67
Naïve bayes	0.64	0.74	0.41	0.55	0.58	**0.64**	0.48	0.59	0.66	0.79

metric	
without Boruta	with Boruta

NANYANG TECHNOLOGICAL UNIVERSITY

Boruta + Local modeling

balanced metric score	
without Boruta	with Boruta

Comparison of balanced evaluation metrics for the local modeling										
	accuracy		precision		recall		F1 score		AUC	
RF	0.72	0.79	0.44	0.78	0.15	0.35	0.20	0.46	0.57	0.83
SVM (rbf)	0.76	**0.85**	0.64	**0.80**	0.35	0.66	0.42	**0.71**	0.72	**0.89**
SVM (linear)	0.73	0.82	0.52	0.70	0.36	0.68	0.41	0.68	0.73	0.85
LG	0.75	0.83	0.62	0.73	0.35	0.65	0.42	0.68	0.74	0.86
KNN	0.72	0.78	0.33	0.71	0.10	0.33	0.14	0.41	0.54	0.70
Naïve bayes	0.66	0.82	0.40	0.64	0.47	**0.72**	0.42	0.68	0.63	0.84

Balanced metric score = $\sum$(class weight $*$ metric score)

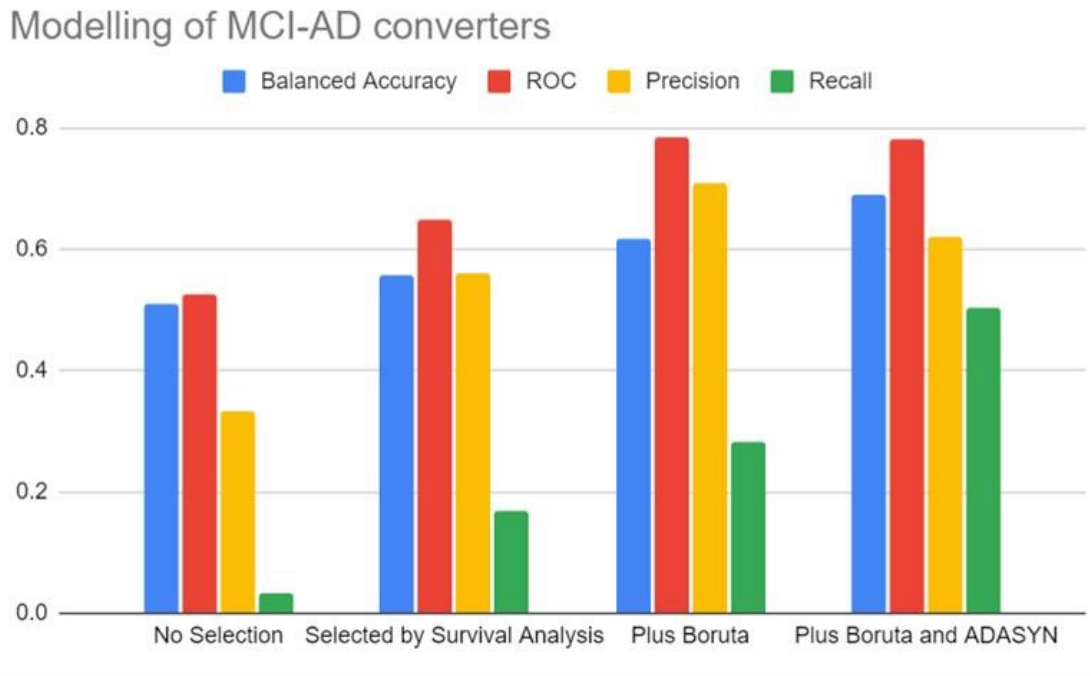
NANYANG TECHNOLOGICAL UNIVERSITY

Comparison of optimal models

Modeling type	optimal model	accuracy	precision	recall	F1-score	AUC
Base modeling	SVM (rbf)	0.83	**0.89**	0.49	0.62	0.86
	SVM (linear)	0.81	0.7	0.61	0.64	0.86
	LG	0.74	0.57	0.53	0.54	0.8
	NB	0.74	0.55	0.64	0.59	0.79
Local modeling	SVM (rbf)	**0.85**	0.8	0.66	**0.71**	**0.89**
	SVM (linear)	0.82	0.7	0.68	0.68	0.85
	LG	0.83	0.73	0.65	0.68	0.86
	NB	0.82	0.64	**0.72**	0.68	0.84

NANYANG TECHNOLOGICAL UNIVERSITY

Work to be done



Apply oversampling strategy (SMOTE):

- Boruta + SMOTE + baseline modeling
- SMOTE + Local modeling
- Boruta + SMOTE + Local modeling

NANYANG TECHNOLOGICAL UNIVERSITY

Appendix: Comparison of evaluation metrics for local modeling

	APOE4	accuracy		precision		recall		F1 score		AUC	
RF	0	0.75	0.79	0.47	0.80	0.08	0.26	0.14	0.39	0.57	0.79
	1	0.66	0.79	0.46	0.86	0.29	0.54	0.35	0.63	0.67	0.87
	2	0.77	0.81	0	0.20	0	0.20	0	0.20	0	0.93
SVM (rbf)	0	0.78	0.86	0.71	0.77	0.29	0.64	0.38	0.69	0.68	0.89
	1	0.73	0.80	0.65	0.80	0.53	0.63	0.58	0.69	0.77	0.88
	2	0.77	1.00	0	1.00	0	1.00	0	1.00	0.77	0.99
SVM (linear)	0	0.75	0.81	0.54	0.66	0.30	0.62	0.37	0.63	0.69	0.85
	1	0.70	0.79	0.59	0.71	0.53	0.71	0.56	0.71	0.78	0.83
	2	0.72	1.00	0	1.00	0	1.00	0	1.00	0.88	0.99
LG	0	0.77	0.82	0.69	0.67	0.28	0.61	0.38	0.63	0.70	0.85
	1	0.71	0.82	0.61	0.83	0.53	0.68	0.57	0.73	0.78	0.84
	2	0.72	0.95	0	0.80	0	0.80	0	0.80	0.88	0.99
KNN	0	0.74	0.77	0.15	0.67	0.04	0.14	0.06	0.23	0.53	0.63
	1	0.66	0.75	0.67	0.72	0.17	0.53	0.27	0.60	0.54	0.77
	2	0.81	1.00	0.20	1.00	0.20	1.00	0.20	1.00	0.65	1.00
Naïve bayes	0	0.64	0.82	0.35	0.64	0.43	0.74	0.37	0.68	0.60	0.84
	1	0.66	0.81	0.55	0.74	0.63	0.78	0.58	0.76	0.70	0.84
	2	0.77	0.82	0	0.20	0	0.20	0	0.20	0.5	0.86

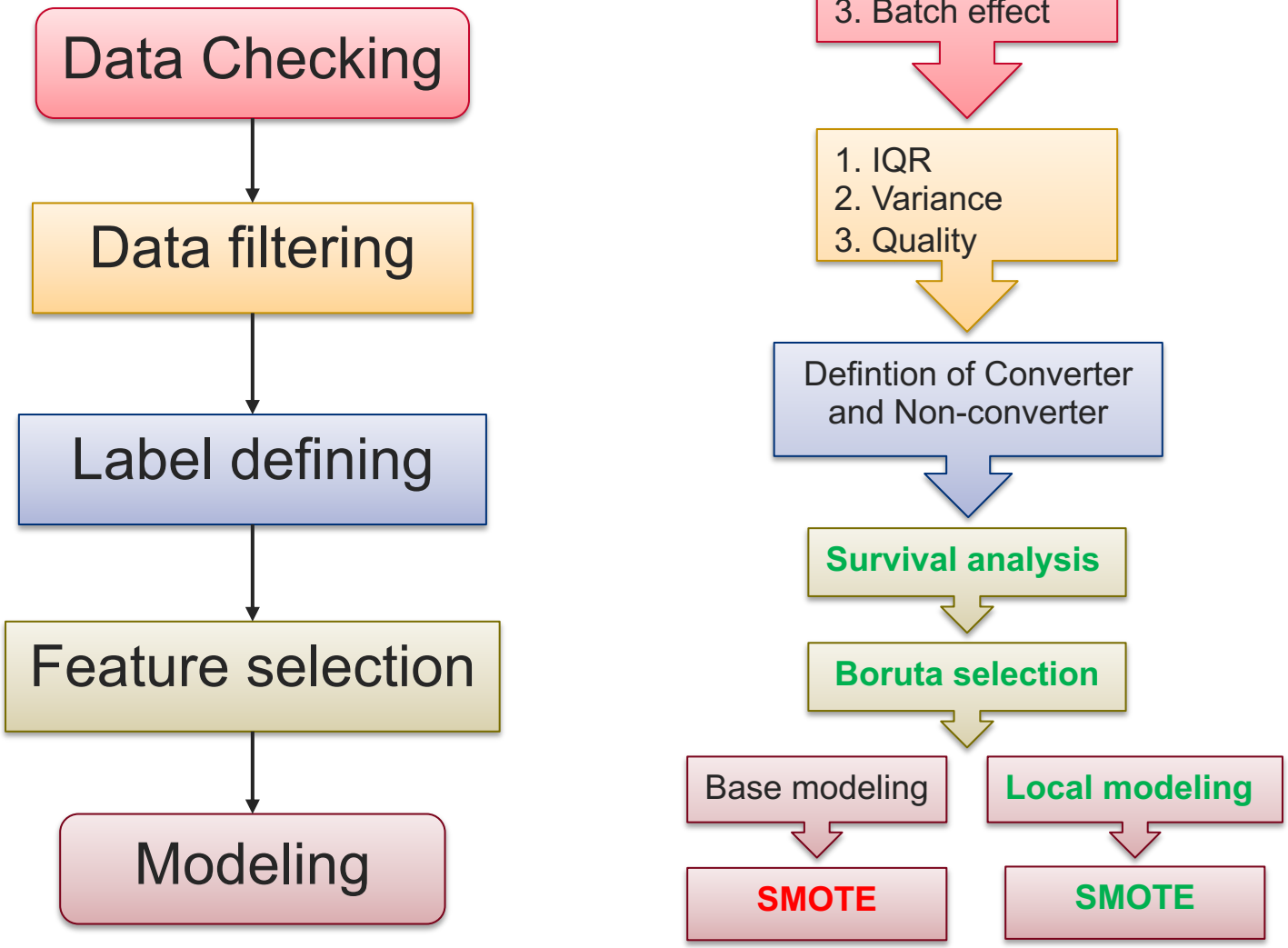
metric	
without Boruta	with Boruta

NANYANG TECHNOLOGICAL UNIVERSITY

Weekly update (Feb/08/2023)

- Story Line
- Exchange the accuracy with AUC as criteria of hyperparameter tuning for RF and SVC
- Apply oversampling strategy after Boruta gene selection
- Methods have been provided: **SMOTE**, **ADASYN**, Borderline SMOTE, SVMSMOTE, KMeansSMOTE, RandomOverSampler, SMOTEENN, SMOTETomek
- Work to be done

Story line



Oversampling

Shape of dataset				
class label	**All samples**	**APOE4 = 0**	**APOE4 = 1**	**APOE4 = 2**
1	95	49	41	5
0	234	145	72	17
Sum	329	194	113	22

SMOTE

Shape of dataset				
class label	**All samples**	**APOE4 = 0**	**APOE4 = 1**	**APOE4 = 2**
1	234	145	72	17
0	234	145	72	17
Sum	468	290	144	34

SMOTE VS. without SMOTE (Boruta + base modeling)

Comparison of evaluation metrics for the optimal base models										
	accuracy		precision		recall		F1 score		AUC	
RF	0.76	0.87	0.83	0.90	0.24	0.85	0.36	0.87	0.79	0.93
SVM (rbf)	0.83	0.90	0.89	0.86	0.49	0.95	0.62	0.90	0.86	**0.97**
SVM (linear)	0.81	**0.92**	0.70	**0.91**	0.61	0.92	0.64	**0.92**	0.86	**0.97**
LogReg	0.74	0.90	0.57	0.86	0.53	0.95	0.54	0.90	0.80	0.96
KNN	0.73	0.63	0.72	0.58	0.16	**0.98**	0.25	0.72	0.67	0.63
Naïve bayes	0.74	0.76	0.55	0.75	0.64	0.80	0.59	0.77	0.79	0.83

metric	
without SMOTE	with SMOTE

NANYANG TECHNOLOGICAL UNIVERSITY

SMOTE VS. without SMOTE (Boruta + local modeling)

Comparison of evaluation metrics for the optimal local models										
	accuracy		precision		recall		F1 score		AUC	
RF	0.79	0.89	0.78	**0.90**	0.35	0.87	0.46	0.88	0.83	0.95
SVM (rbf)	0.85	**0.90**	0.80	0.88	0.66	0.92	0.71	0.90	0.89	**0.97**
SVM (linear)	0.82	0.88	0.70	0.85	0.68	0.89	0.68	0.87	0.85	**0.97**
LogReg	0.83	**0.90**	0.73	**0.90**	0.65	0.91	0.68	**0.91**	0.86	**0.97**
KNN	0.78	0.70	0.71	0.65	0.33	**0.95**	0.41	0.77	0.70	0.70
Naïve bayes	0.82	0.83	0.64	0.86	0.72	0.80	0.68	0.82	0.84	0.90

Balanced metric score = $\sum$(class weight $*$ metric score)

metric	
without SMOTE	with SMOTE

NANYANG TECHNOLOGICAL UNIVERSITY

Comparison of optimal models

Modeling type	optimal model	accuracy	precision	recall	F1-score	AUC
Base modeling	RF	0.87	0.90	0.85	0.87	0.95
	SVM (rbf)	0.90	0.86	**0.95**	0.90	**0.97**
	SVM (linear)	**0.92**	**0.91**	0.92	**0.92**	**0.97**
	LogReg	0.90	0.86	**0.95**	0.90	**0.97**
	NB	0.76	0.75	0.80	0.77	0.83
Local modeling	RF	0.89	0.90	0.87	0.88	0.95
	SVM (rbf)	0.90	0.88	0.92	0.90	**0.97**
	SVM (linear)	0.88	0.85	0.89	0.87	**0.97**
	LogReg	0.90	0.90	0.91	0.91	**0.97**
	NB	0.83	0.86	0.80	0.82	0.90

Work to be done

- Correct the oversampling methods
- Evaluate the data preprocessing methods
- Try other oversampling strategy(ADASYN)

Appendix: Comparison of evaluation metrics for local modeling(with boruta)

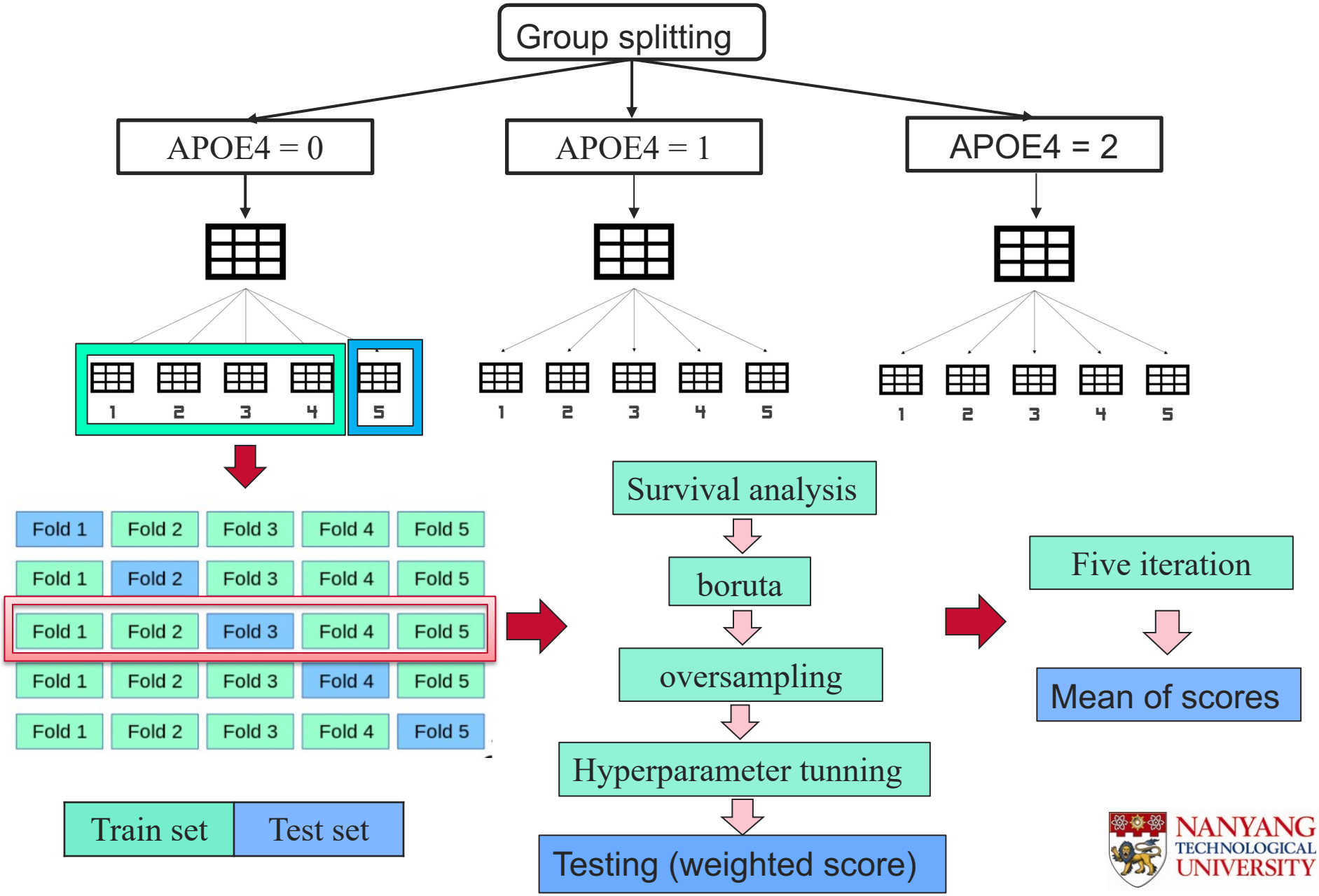
	APOE4	accuracy		precision		recall		F1 score		AUC	
RF	0	0.79	0.90	0.80	0.91	0.26	0.89	0.39	0.90	0.79	0.96
	1	0.79	0.85	0.86	0.87	0.54	0.82	0.63	0.84	0.87	0.91
	2	0.81	0.94	0.20	1.00	0.20	0.87	0.20	0.92	0.93	0.98
SVM (rbf)	0	0.86	0.92	0.77	0.90	0.64	0.94	0.69	0.92	0.89	0.98
	1	0.80	0.83	0.80	0.82	0.63	0.86	0.69	0.84	0.88	0.93
	2	1.00	0.97	1.00	0.96	1.00	1.00	1.00	0.98	0.99	0.99
SVM (linear)	0	0.81	0.92	0.66	0.90	0.62	0.94	0.63	0.92	0.85	0.98
	1	0.79	0.87	0.71	0.87	0.71	0.87	0.71	0.87	0.83	0.94
	2	1.00	0.54	1.00	0.37	1.00	0.60	1.00	0.44	0.99	0.99
LogReg	0	0.82	0.92	0.67	0.92	0.61	0.94	0.63	0.93	0.85	0.98
	1	0.82	0.85	0.83	0.85	0.68	0.86	0.73	0.85	0.84	0.94
	2	0.95	0.94	0.80	0.96	0.80	0.93	0.80	0.94	0.99	0.99
KNN	0	0.77	0.66	0.67	0.60	0.14	0.95	0.23	0.73	0.63	0.65
	1	0.75	0.72	0.72	0.66	0.53	0.95	0.60	0.78	0.77	0.73
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Naïve bayes	0	0.82	0.82	0.64	0.86	0.74	0.77	0.68	0.81	0.84	0.88
	1	0.81	0.82	0.74	0.83	0.78	0.82	0.76	0.82	0.84	0.91
	2	0.82	0.97	0.20	1.00	0.20	0.93	0.20	0.96	0.86	0.99

metric	
without SMOTE	with SMOTE

Feb 17 2023 weekly update

- Correction of errors
- Correction the pipeline and evaluation

Modified pipeline and evaluation methods



NANYANG TECHNOLOGICAL UNIVERSITY

Survival analysis + boruta + SMOTE + base modeling

	Accuracy	Balanced accuracy	precision	recall	F1-score	AUC
RF	0.62	0.51	0.31	0.24	0.26	0.51
SVC(rbf)	0.57	0.54	0.35	0.47	0.37	0.54
SVC(linear)	0.57	0.54	0.35	0.47	0.37	0.54

Distribution of class label (training set)						
label	**Fold 0**	**Fold 1**	**Fold 2**	**Fold3**	**Fold 4**	**percentage**
0	187	187	187	187	188	71%
1	76	76	76	76	76	29%

Distribution of class label (testing set)						
label	**Fold 0**	**Fold 1**	**Fold 2**	**Fold3**	**Fold 4**	**percentage**
0	47	47	47	47	46	71%
1	19	19	19	19	19	29%

Mar 01 Weekly Update

- Remove the data leakage
- Generate the result for whole pipeline

Result: Base modeling

	Accuracy	Balanced accuracy	precision	recall	F1-score	AUC
RF	**0.66**	0.49	0.20	0.08	0.11	0.49
SVC(rbf)	0.52	**0.54**	0.34	0.59	**0.41**	0.54
SVC(linear)	0.58	0.44	0.18	0.13	0.15	0.52
RF + SMOTE	0.64	0.51	0.32	0.22	0.24	0.52
SVC (rbf) + SMOTE	0.57	**0.54**	**0.35**	0.47	0.37	0.54
SVC (linear) + SMOTE	0.62	0.52	0.32	0.31	0.31	0.50
RF + ADASYN	0.63	**0.54**	**0.35**	0.29	0.31	**0.55**
SVC (rbf) + ADASYN	0.38	0.51	0.23	**0.8**	0.36	0.51
SVC (linear) + ADASYN	0.61	0.52	0.32	0.31	0.31	0.51

Result: Local modeling

	Accuracy	Balanced accuracy	precision	recall	F1-score	AUC
RF	0.68	0.49	0.17	0.07	0.10	**0.50**
SVC(rbf)	0.62	0.5	0.08	0.26	0.12	0.47
SVC(linear)	0.66	0.50	0.32	0.19	0.23	0.44
RF + SMOTE	0.66	0.50	0.31	0.20	0.21	0.49
SVC (rbf) + SMOTE	0.63	0.50	0.30	0.27	**0.27**	0.48
SVC (linear) + SMOTE	0.63	0.49	0.29	0.25	0.25	0.48
RF + ADASYN	**0.69**	**0.52**	**0.34**	0.22	0.24	0.48
SVC (rbf) + ADASYN	0.56	0.49	0.22	**0.39**	0.26	0.43
SVC (linear) + ADASYN	0.65	0.50	0.30	0.24	0.25	0.51

Base modeling vs. Local modeling

	Accuracy		Balanced accuracy		precision		recall		F1-score		AUC	
RF	0.66	0.68	0.49	0.49	0.20	0.17	0.08	0.07	0.11	0.10	0.49	0.50
SVC(rbf)	0.52	0.62	**0.54**	0.5	0.34	0.08	0.59	0.26	**0.41**	0.12	0.54	0.47
SVC (linear)	0.58	0.66	0.44	0.50	0.18	0.32	0.13	0.19	0.15	0.23	0.52	0.44
RF + SMOTE	0.64	0.66	0.51	0.50	0.32	0.31	0.22	0.20	0.24	0.21	0.52	0.49
SVC (rbf) + SMOTE	0.57	0.63	**0.54**	0.50	**0.35**	0.30	0.47	0.27	0.37	0.27	0.54	0.48
SVC (linear) + SMOTE	0.62	0.63	0.52	0.49	0.32	0.29	0.31	0.25	0.31	0.25	0.50	0.48
RF + ADASYN	0.63	**0.69**	**0.54**	0.52	**0.35**	0.34	0.29	0.22	0.31	0.24	**0.55**	0.48
SVC (rbf) + ADASYN	0.38	0.56	0.51	0.49	0.23	0.22	**0.8**	0.39	0.36	0.26	0.51	0.43
SVC (linear) + ADASYN	0.61	0.65	0.52	0.50	0.32	0.30	0.31	0.24	0.31	0.25	0.51	0.51

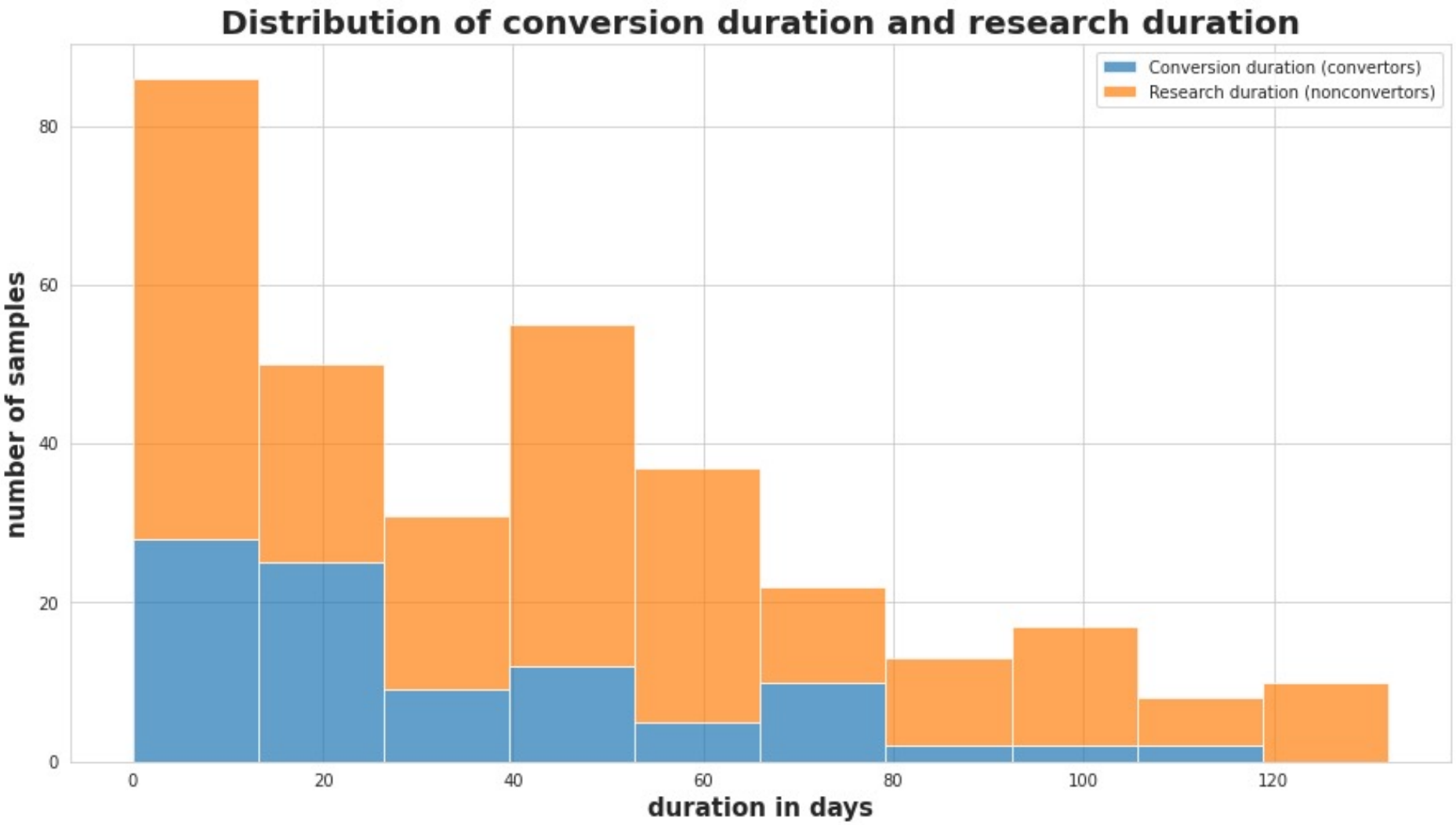
Evaluation metric	
Base modeling	Local modeling

NANYANG TECHNOLOGICAL UNIVERSITY

Mar 10 Weekly Update

- Applied Scikit survival on dataset
- Predict the risk of conversion
- Predict the conversion time using regression

Distribution of conversion duration and research duration



NANYANG TECHNOLOGICAL UNIVERSITY

Concordance index

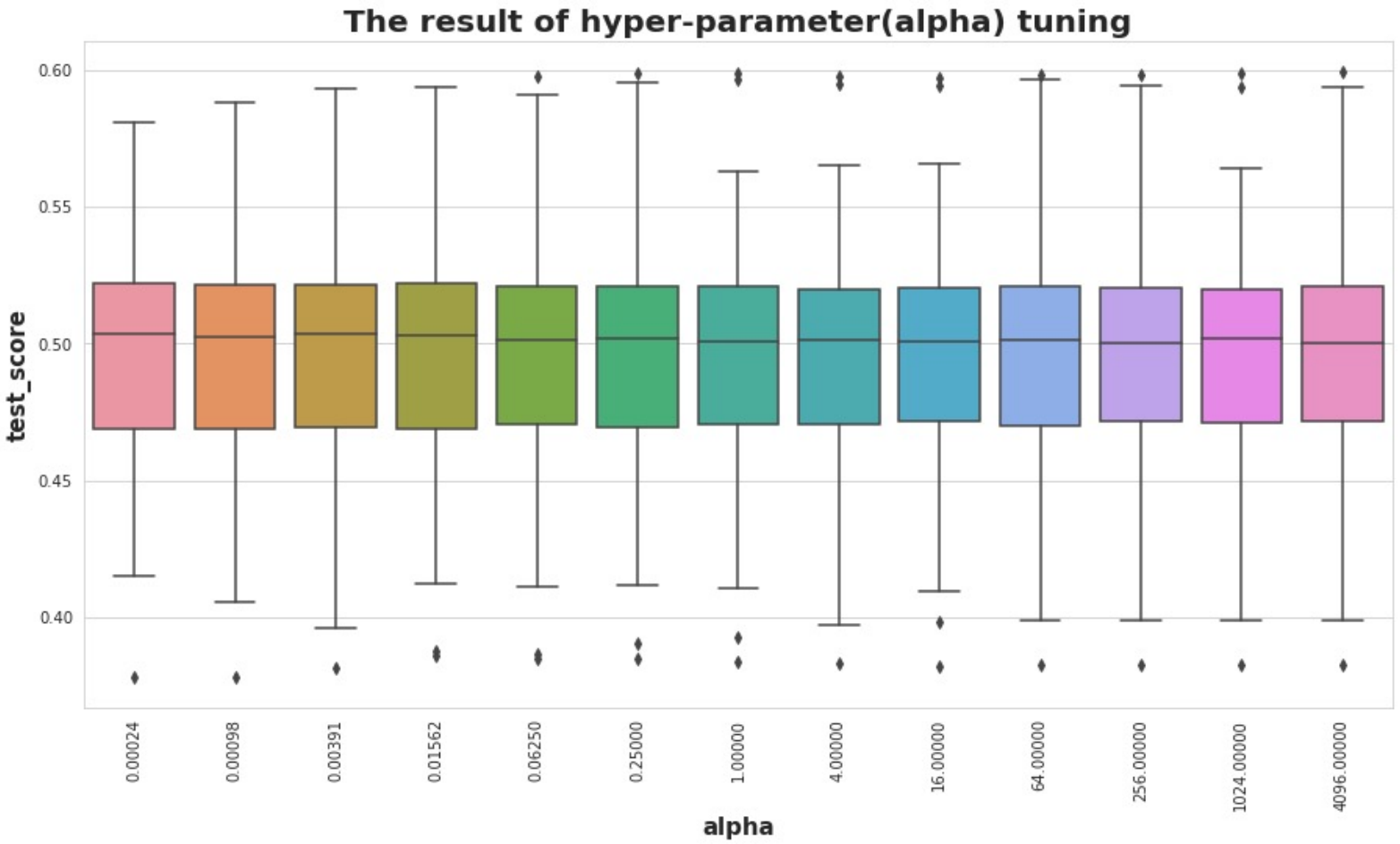
- It measures the ability of a model to correctly rank the survival times of pairs of individuals.

- The C-index ranges from 0.5 (indicating a model with no predictive ability) to 1.0 (indicating a model with perfect predictive ability).

- alpha (hyper-parameter): Weight of penalizing the squared hinge loss in the objective function.

- alpha determines the amount of regularization to apply: A smaller value increases the amount of regularization and a higher value reduces the amount of regularization.

Result of conversion risk prediction



Model	Optimal C-index	Optimal alpha
Linear Survival SVM	0.498	0.00024

NANYANG TECHNOLOGICAL UNIVERSITY

Result of conversion time prediction

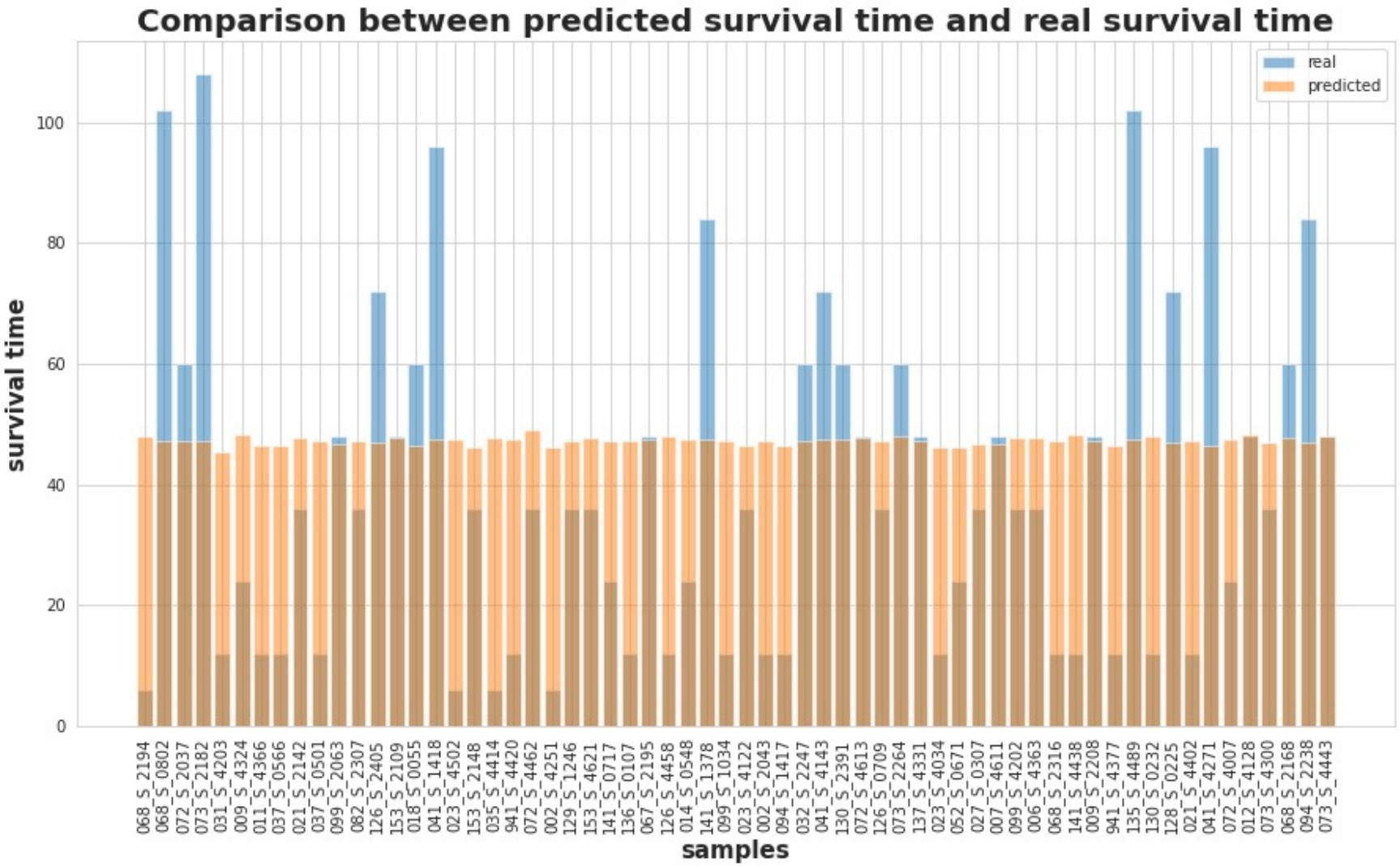
Table 1. Dataset splitting

	Sample size	percentage
training	244	80%
testing	62	20%

Table 2. Evaluation scores of prediction

	MSE (days)	RMSE (days)
Scores	818	28

Conversion time: real vs. predicted



NANYANG TECHNOLOGICAL UNIVERSITY