RandomForest VS GradientBoosting

Eunju Lee

Department of Statistics
Pusan National University

September 15, 2021

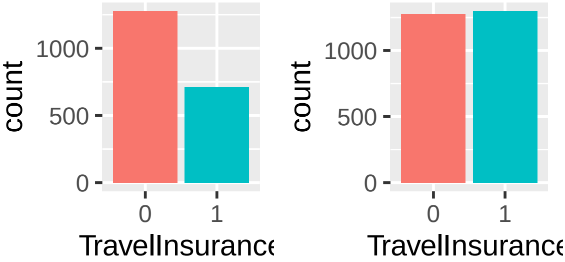
1. Travel Insurance

Data Description

- Age
- Employment Type (1:Government 0:Private/Self Employed)
- GraduateOrNot (1:Yes 0:No)
- Annual Income
- FamilyMembers
- ChronicDisease (1:Yes 0:No)
- FrequentFlyer (1:Yes 0:No)
- EverTravelledAbroad (1:Yes 0:No)
- TravelInsurance (1:Yes 0:No)

2. Preprocessing & Modeling Preparation

- **NA** : 0 rows
- **Y class rate** : Up sampling

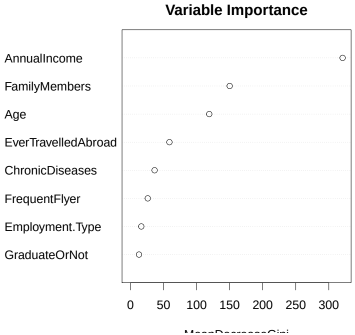


- **train vs test** : train set 70%, test set 30%
- **Hyper parameter** : Grid Search (Repeated Cross Validation)

3. RandomForest

- modeling

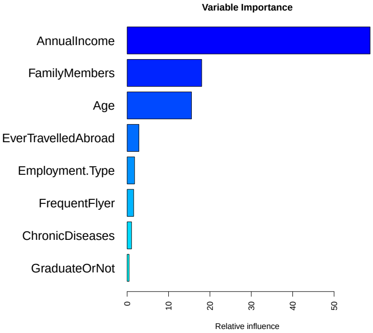
	$0(\hat{Y})$	$1(\hat{Y})$	class error
0(Y)	780	114	0.13
1(Y)	140	770	0.15



4. GBM

- modeling

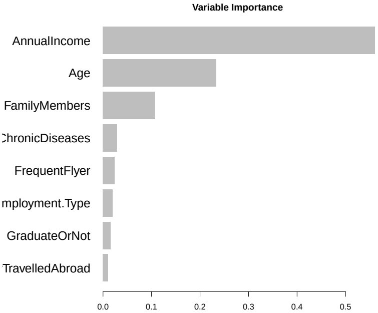
	$0(\hat{Y})$	$1(\hat{Y})$	class error
0(Y)	873	21	0.02
1(Y)	340	570	0.37



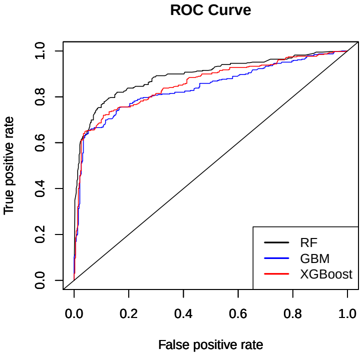
5. XGBoost

- modeling

	$0(\hat{Y})$	$1(\hat{Y})$	class error
0(Y)	831	63	0.07
1(Y)	193	717	0.21



6. Comparing models



Model	AUC
RF	0.89
GBM	0.84
XGBoost	0.86

1. Heart Risk

Data Description

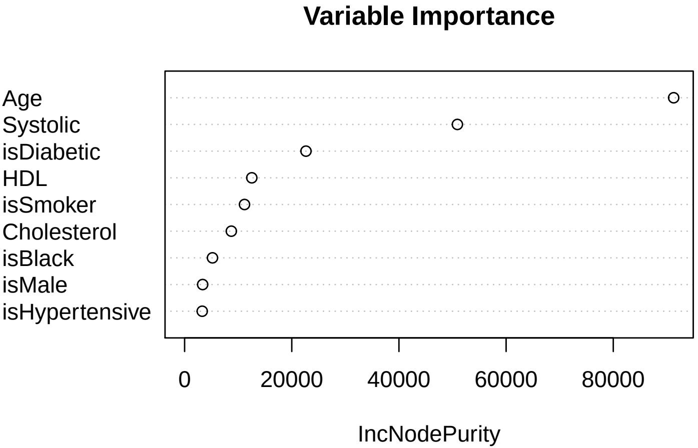
- isMale (1:Male 0:Female)
- isBlack (1:Black 0:Not)
- isSmoker (1:Smoker 0:Non-smoker)
- isDiabetic (1:Diabetic 0:Normal)
- isHypertensive (1:Yes 0:No)
- Age
- Systolic (Maximum Blood Pressure)
- Cholesterol
- HDL
- Risk(%)

2. Preprocessing & Modeling Prearation

- **NA** : 0 rows
- **train vs test** : train set 70%, test set 30%
- **Hyper parameter** : Grid Search (Repeated Cross Vaslidation)

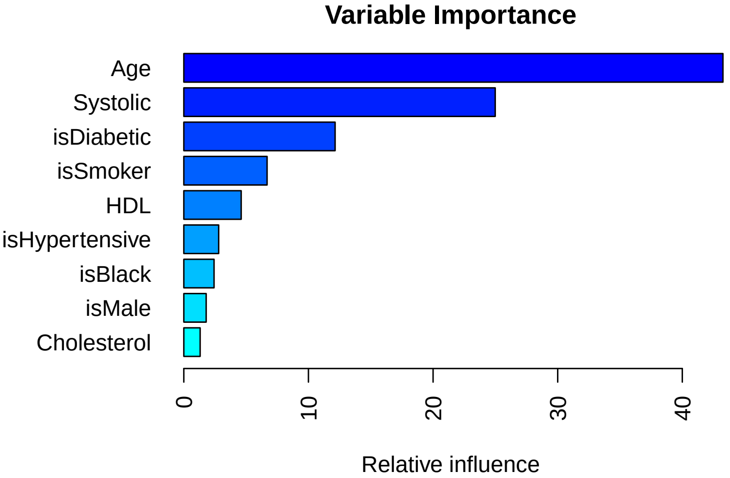
3. RandomForest

- modeling - RMSE : 6.31



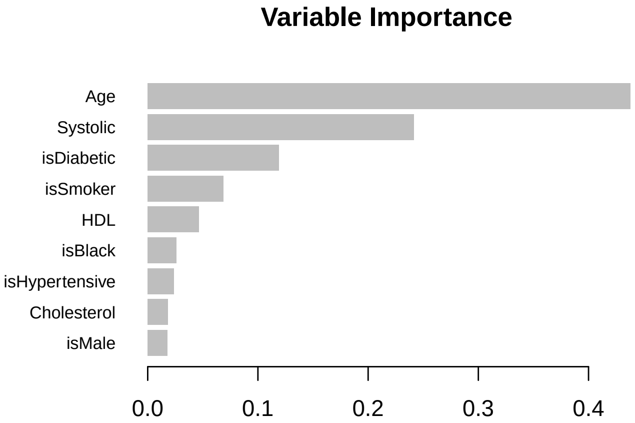
4. GBM

- modeling - RMSE : 3.80



5. XGBoost

- modeling - RMSE : 1.67



6. Comparing Models

Model	RMSE(train)	RMSE	R^2
RF	6.31	6.21	0.79
GBM	3.80	4.94	0.89
XGBoost	1.67	4.87	0.90

Comparing 2 Data

- Travel Insurance : Classification(AUC)
- Heart Risk : Regression(R^2)

	Travel Insurance	Heart Risk
RF	0.89	0.79
GBM	0.84	0.89
XGBoost	0.86	0.90