Lifestyle Risk Factors for Suicide

Dahee Lee, Lu Ding, Merita Seferi

Overview of Predictors and Outcome

Outcome: Suicide in Adults

Lifestyle-Related Predictors:

- ❖Substances use: Tobacco, Marijuana, Cocaine, and Alcohol consumption
- ❖Employment
- ❖Years of education
- ❖Age
- ❖Gender
- ❖Diet: Decreased appetite, and increased appetite
- ❖Mental Health: History of Major Depressive Episode (MDE)

Importance of the Model

- Suicide is the10th leading cause of death in the US
- On average, there are 121 suicides per day
- Suicide costs the US 51 $ Billion annually
- Stigma surrounding suicide leads to underreporting, and data collection methods critical to suicide prevention need to be improved

Importance of the Model

- Evidence of other lifestyle predictors in literature:
- We recognize that people with healthier lifestyles are likely to have better health outcomes in general, but would like to see <u>which</u> lifestyle habits are most important predictors of suicide in particular

Overview of the dataset

- National Survey on Drug Use and Health, 2014
- The survey is designed to provide information such as use of illicit drugs, alcohol, and tobacco, mental health and general demographics among members of United States households aged 12 and older. (155,271)
- The survey includes questions concerning treatment for both substance abuse and mental health-related disorders.

Overview of the dataset

- Respondents were asked about personal and family income sources and amounts, health care access and coverage, neighborhood environment, illegal activities and arrest record.

- Background information include gender, race, age, ethnicity, marital status, educational level, job status, veteran status, and current household composition.

Overview of the dataset

Advantages of our dataset:

- Very large and diverse sample
- Had the outcome and predictors we were interested in

Dealing with Missing Values

- Raw data: 55271 observations in total
- 24286 missing values in the outcome variable
- Removed all the missing values in “suicide”
- Adults only: 4309 available observations

Dealing with Missing Values

➢ Missingness for all the variables that we chose

	type	missing	method	model
age	ordered-categorical	0	<NA>	<NA>
sex	binary	0	<NA>	<NA>
tobacco	binary	0	<NA>	<NA>
alcohol	binary	0	<NA>	<NA>
marijuana	binary	2	ppd	logit
cocaine	binary	1	ppd	logit
job	unordered-categorical	312	ppd	mlogit
education	unordered-categorical	0	<NA>	<NA>
eatingSmall	binary	25	ppd	logit
eatingLarge	binary	2612	ppd	logit
suiThink	binary	0	<NA>	<NA>

Dealing with Missing Values

- Created TableOne:

No dependency -- MCAR

Only for “eatingLarge” →

	Stratified by missing FALSE	TRUE	p	test
n	1697	2612		
age (mean (sd))	3.96 (0.75)	3.93 (0.75)	0.125	
sex = m (%)	610 (35.9)	909 (34.8)	0.462	
tobacco = y (%)	1304 (76.8)	2043 (78.2)	0.307	
alcohol = y (%)	1606 (94.6)	2474 (94.7)	0.965	
marijuana = y (%)	1114 (65.7)	1775 (68.0)	0.125	
cocaine = y (%)	435 (25.6)	755 (28.9)	0.021	
job (%)			0.025	
disabled	200 (12.7)	402 (16.6)		
full	866 (54.9)	1254 (51.8)		
house	111 (7.0)	154 (6.4)		
part	217 (13.8)	302 (12.5)		
retired	95 (6.0)	162 (6.7)		
school	23 (1.5)	34 (1.4)		
unemployed	66 (4.2)	111 (4.6)		
education (%)			<0.001	
<high	139 (8.2)	299 (11.4)		
gradutate	664 (39.1)	887 (34.0)		
high	390 (23.0)	639 (24.5)		
some	504 (29.7)	787 (30.1)		
eatingSmall = y (%)	0 (0.0)	2590 (99.8)	<0.001	
eatingLarge = y (%)	936 (55.2)	0 (NaN)	NaN	
suiThink = y (%)	786 (46.3)	1359 (52.0)	<0.001	
missing = TRUE (%)	0 (0.0)	2612 (100.0)	<0.001	

Dealing with Missing Values

Raw data

 # Total observations = 55271

Available outcome variable (only adult)

4309

→ Naive-Bayes

In R('classif.naiveBayes') implementation, it takes NAs, as it will automatically remove all the NAs.

Removing every missing variable among predictive variable

1570

→ Random Forest

Machine learning Algorithm

- Naive-Bayes :
  - Easy to implement, very efficient
  - There are many missing values in "eatingLarge", and Naive-Bayes can handle the missingness by itself

- NEW -- Random Forest:
  - Better than Decision Tree, Bagging and Boosting, because it randomly picks subsets (not correlated)

Naive-Bayes

- Naive-Bayes：

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\, P(c)}{P(x)}$$

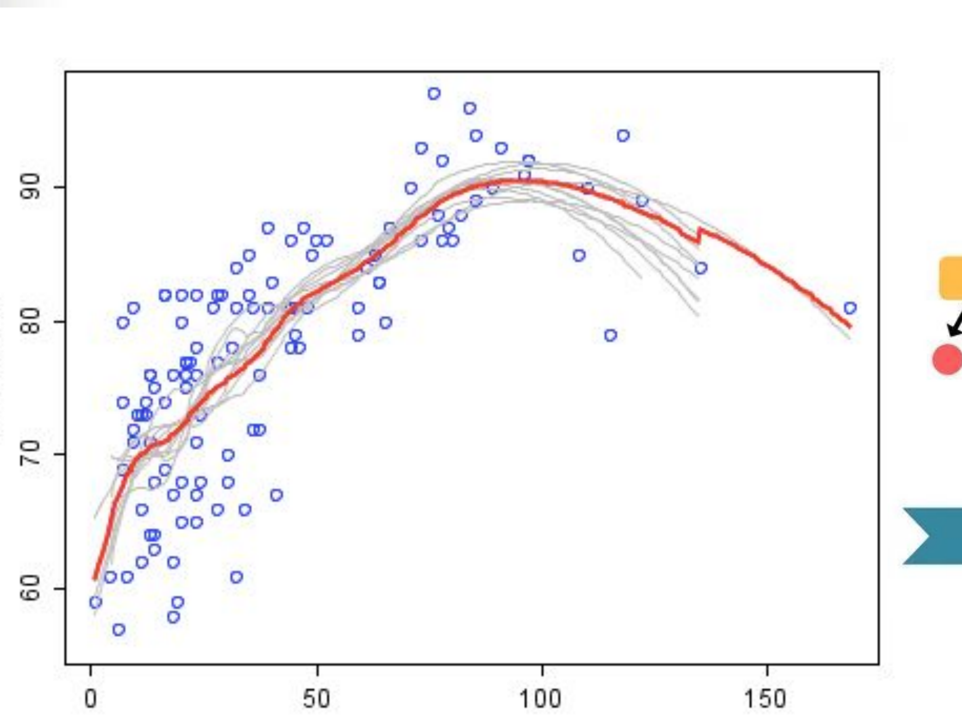
Posterior Probability

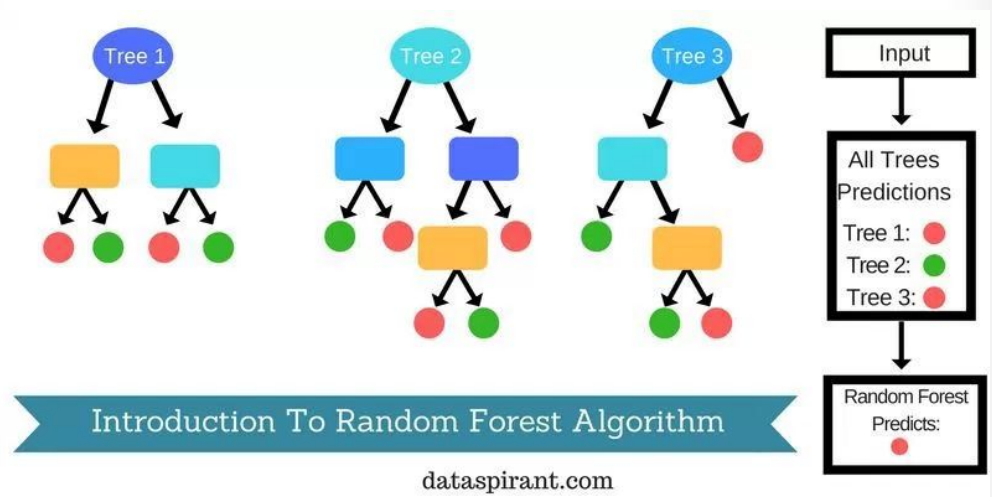
Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Random Forest

- **Random Forest:**
  - An ensemble approach that can also be thought of as a form of nearest neighbor predictor.





Naive-Bayes vs Random Forest

- ➢ Performance

Naive-Bayes

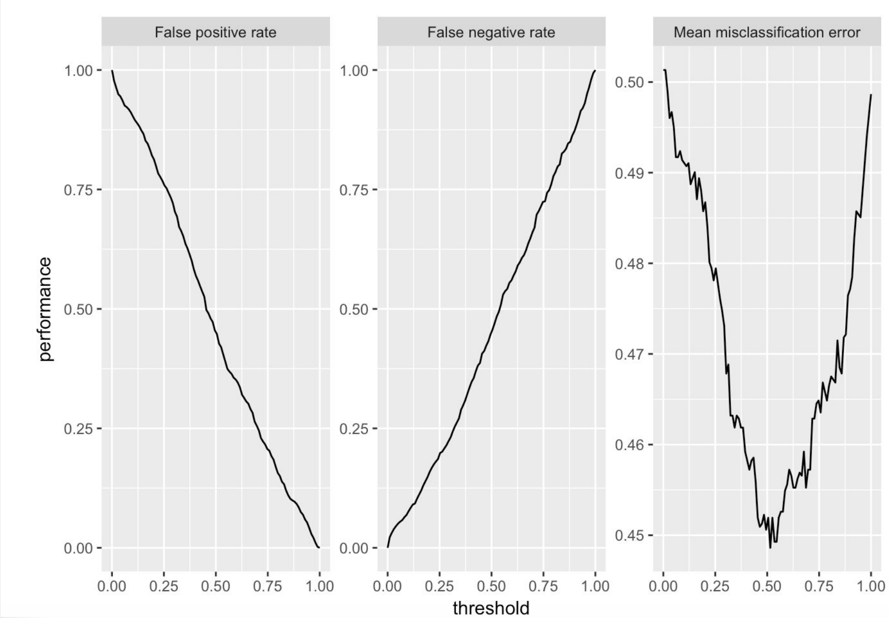
iter	acc	auc	ppv	tpr
1	0.5487078	0.5675272	0.5473888	0.5626243
2	0.5432836	0.5432541	0.5497076	0.5529412
3	0.5542289	0.5715664	0.5450734	0.5295316

RandomForest

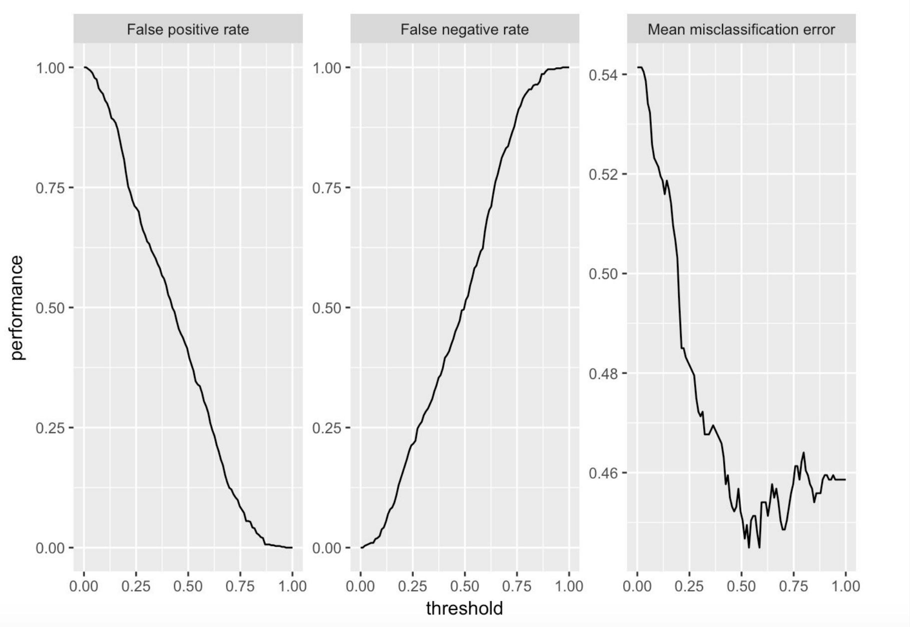
iter	acc	auc	ppv	tpr
1	0.5519126	0.5715815	0.5030303	0.5030303
2	0.5367847	0.5453649	0.4966887	0.4437870
3	0.5546448	0.5553421	0.5204678	0.5235294

Naive-Bayes vs Random Forest

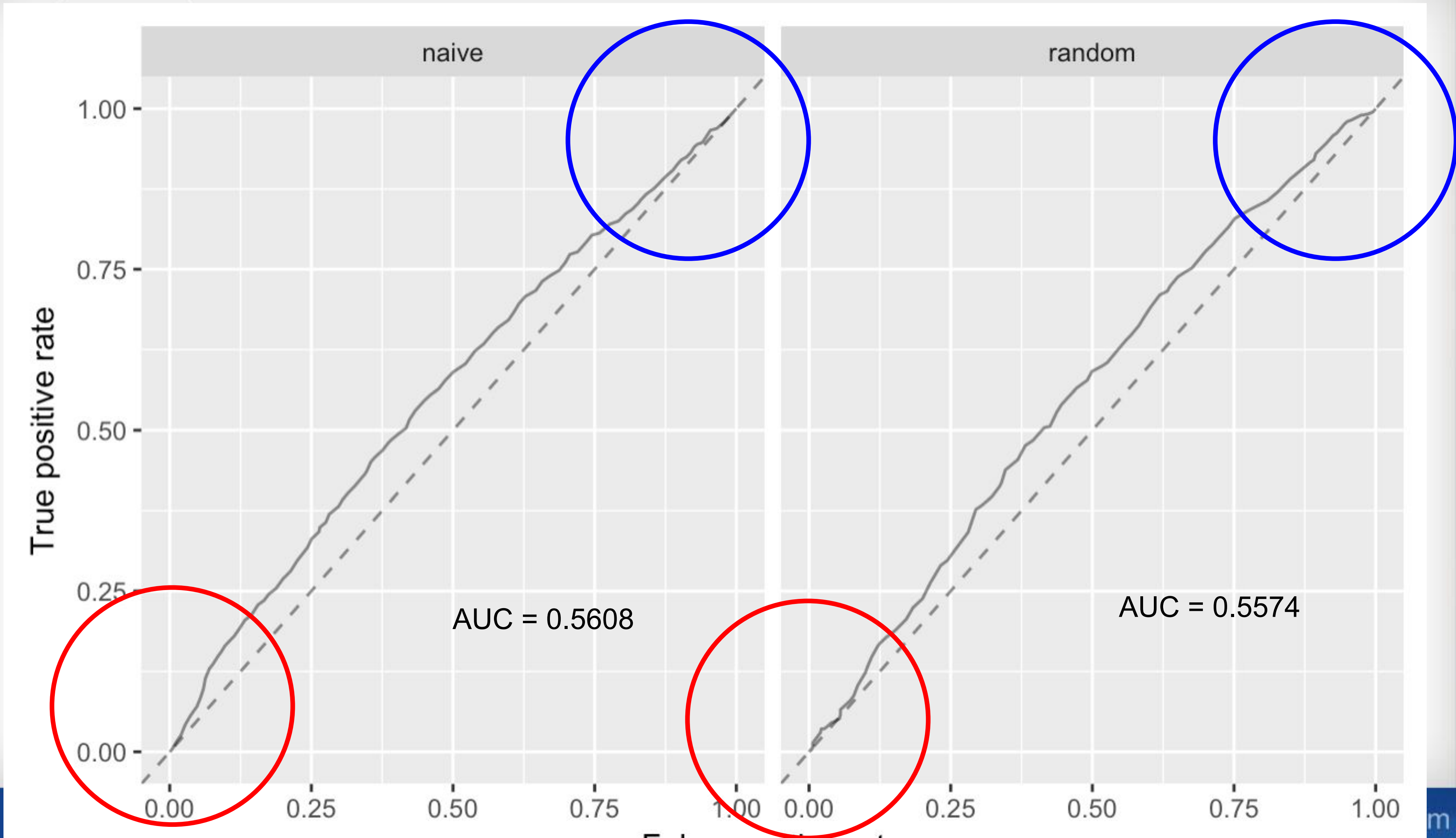
Naive-Bayes



RandomForest



Naive-Bayes vs Random Forest



How to increase accuracy?

- Include another variable (history of seeing doctor for MDE)

- Use Naive-Bayes (because its overall performance is slightly better than random forest)

Additional feature vs Original set

Adding new predictor

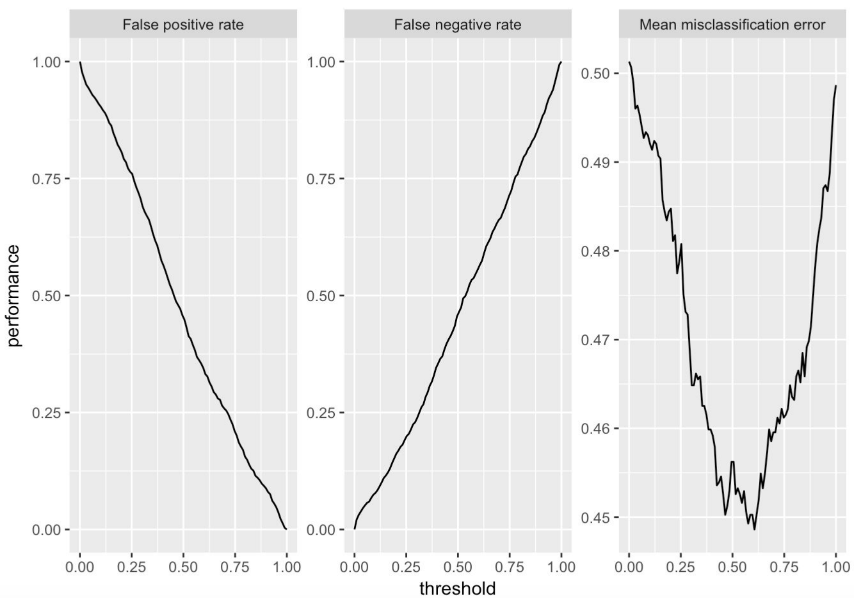
iter	acc	auc	ppv	tpr
1	0.5467197	0.5704086	0.5459883	0.5546720
2	0.5263682	0.5414419	0.5332031	0.5352941
3	0.5552239	0.5721984	0.5466102	0.5254582

Original Naive-Bayes

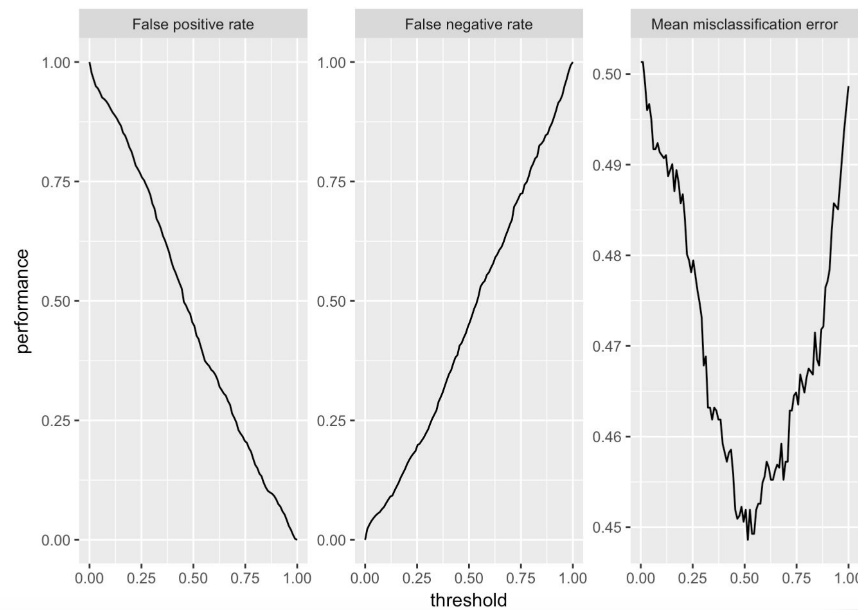
iter	acc	auc	ppv	tpr
1	0.5487078	0.5675272	0.5473888	0.5626243
2	0.5432836	0.5432541	0.5497076	0.5529412
3	0.5542289	0.5715664	0.5450734	0.5295316

Additional feature vs Original set

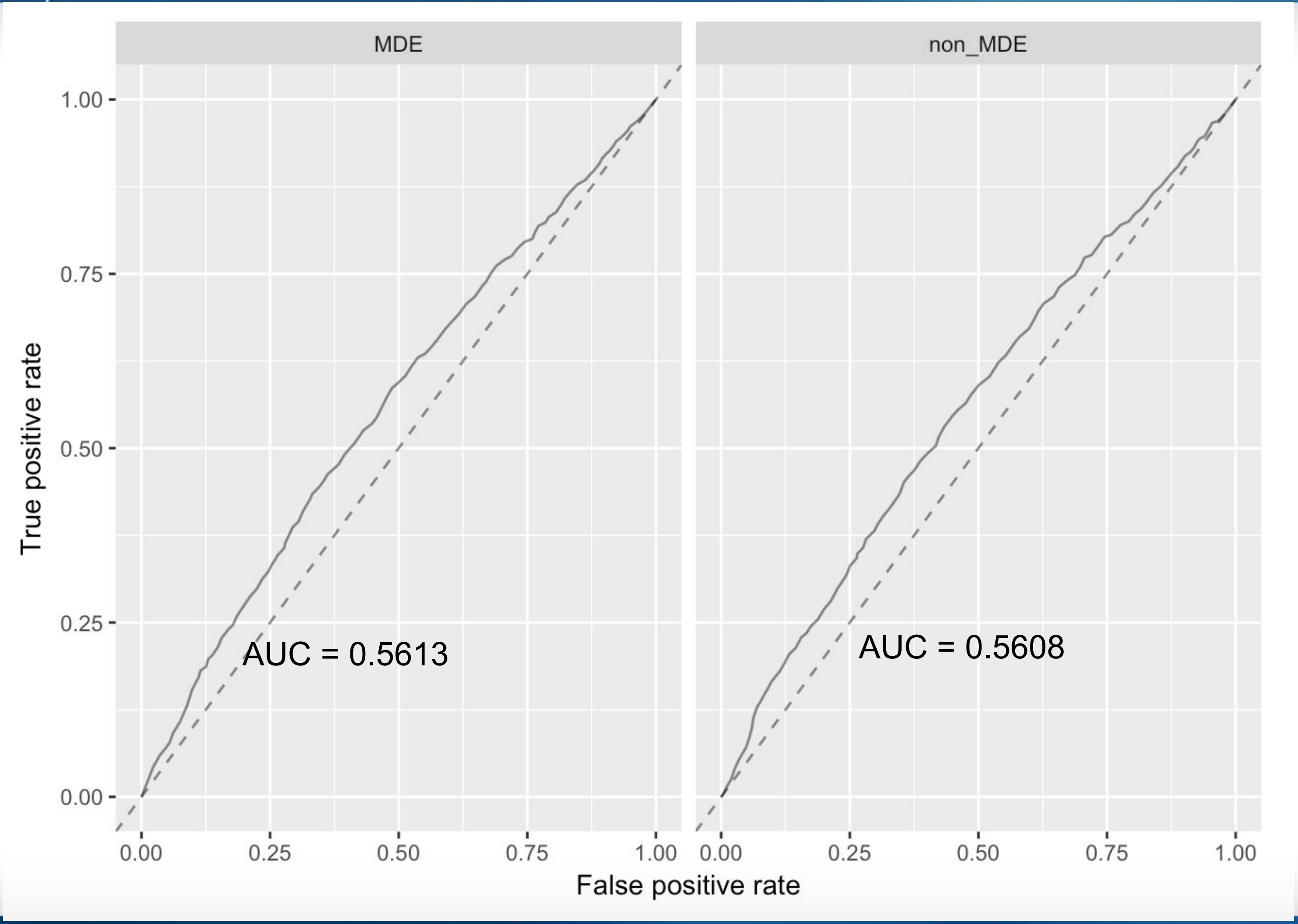
Adding new predictor



Original Naive-bayes



Additional feature vs Original set



Conclusion

- Beneficial for healthcare providers to screen the possibility of suicide
- Naive-Bayes algorithm is slightly better

- Limitation: The accuracy is not high enough.

- Next step to a better model:
  Eliminate some features

Acknowledgements

- Dahee Lee: Algorithms and codes
- Lu Ding: Algorithms and presentation
- Merita Seferi: Overview

Thank you!