USC University of Southern California

Stroke Prediction

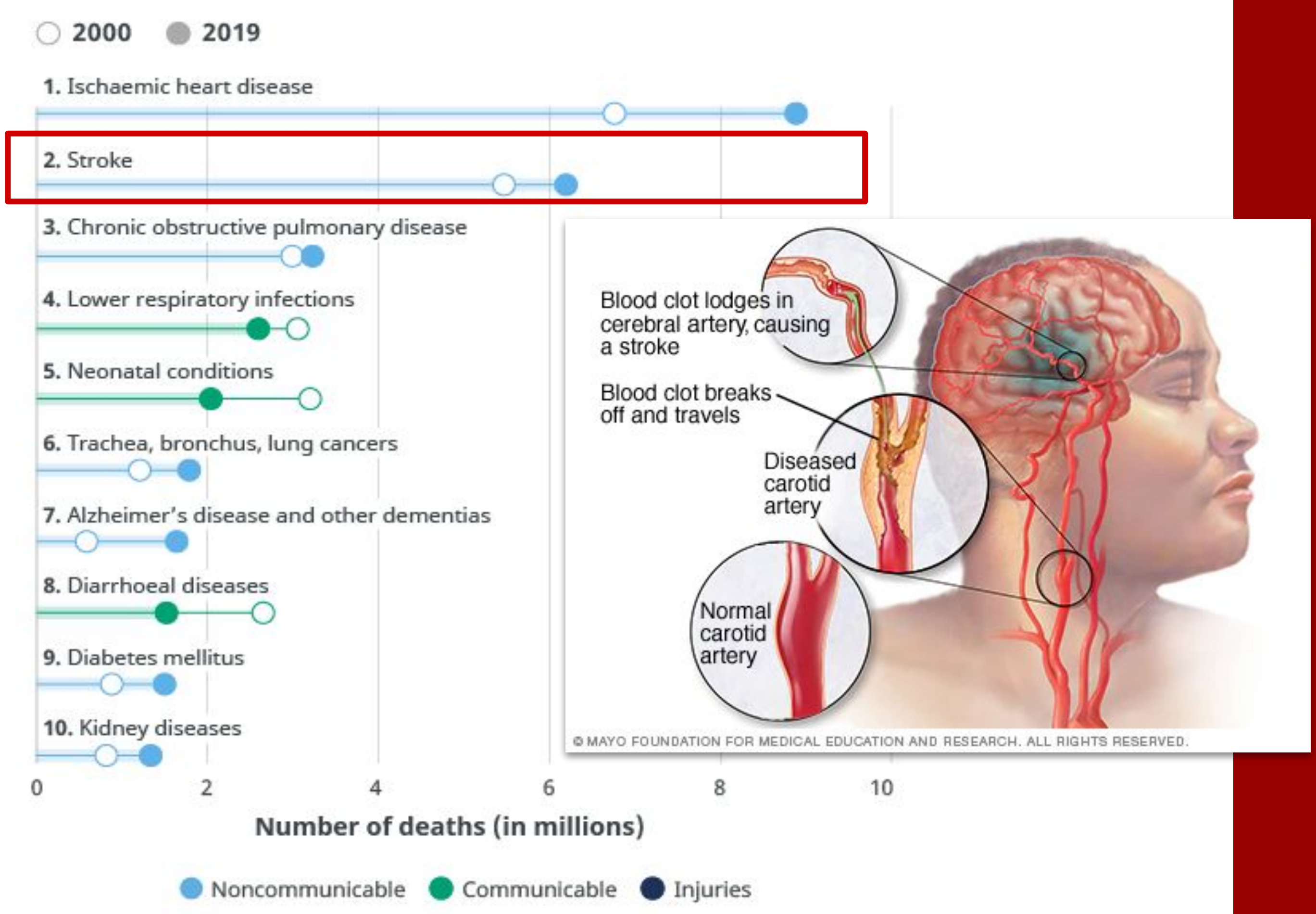
ISE 599 - Introduction to Health Analytics

Group Member & Student ID

Panissara Varin	6462987857
Heyujia Du	5241401889
Zhihao Guo	1958282693

The global ranking of stroke as the second leading cause of death

Leading causes of death globally



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Source: WHO Global Health Estimates.

~15M individuals globally

incidence of stroke affects **every year**

~1/3 death rate

survivors with long-term consequences in **vision, speech impairments, paralysis**

~1/4 stroke survivors experiencing another stroke **within 5 years**

Data from World Health Organization and Center of Disease Control and Prevention

USC

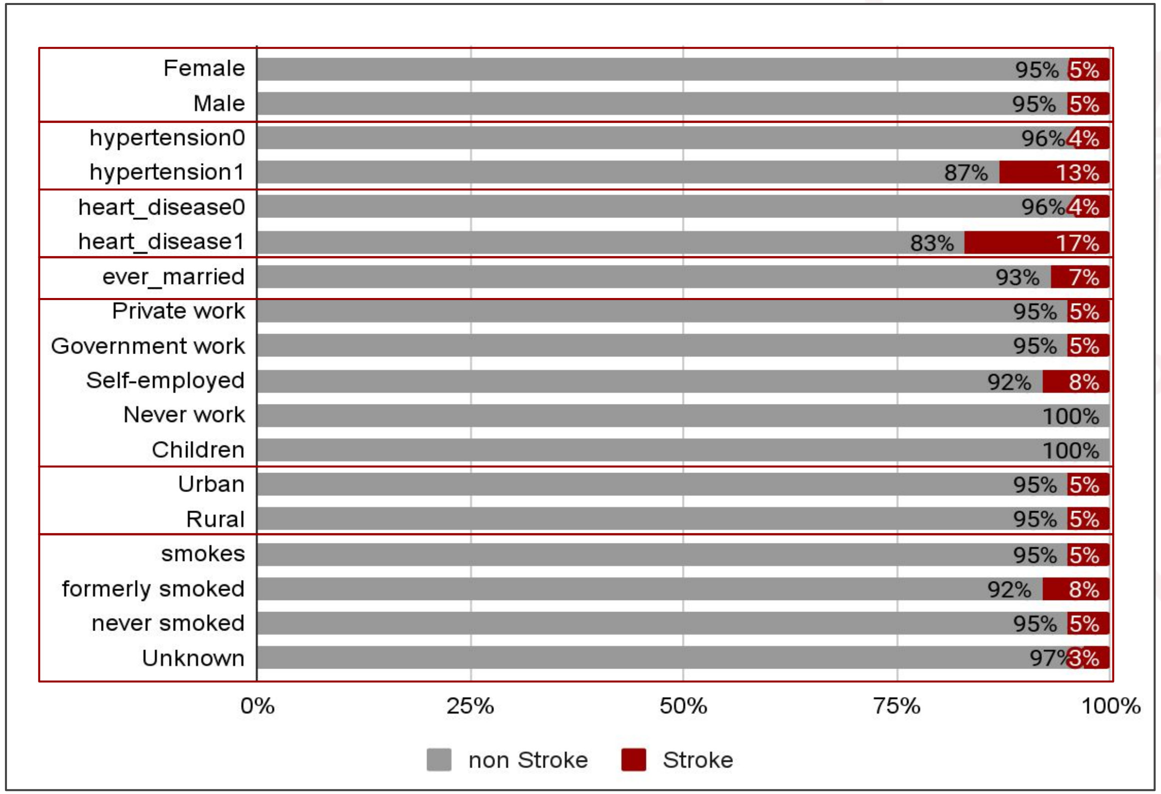
Objective

Identifying **significant contributing factors** that influence stroke occurrence and to **develop models** that can accurately predict a patient's risk of experiencing a stroke

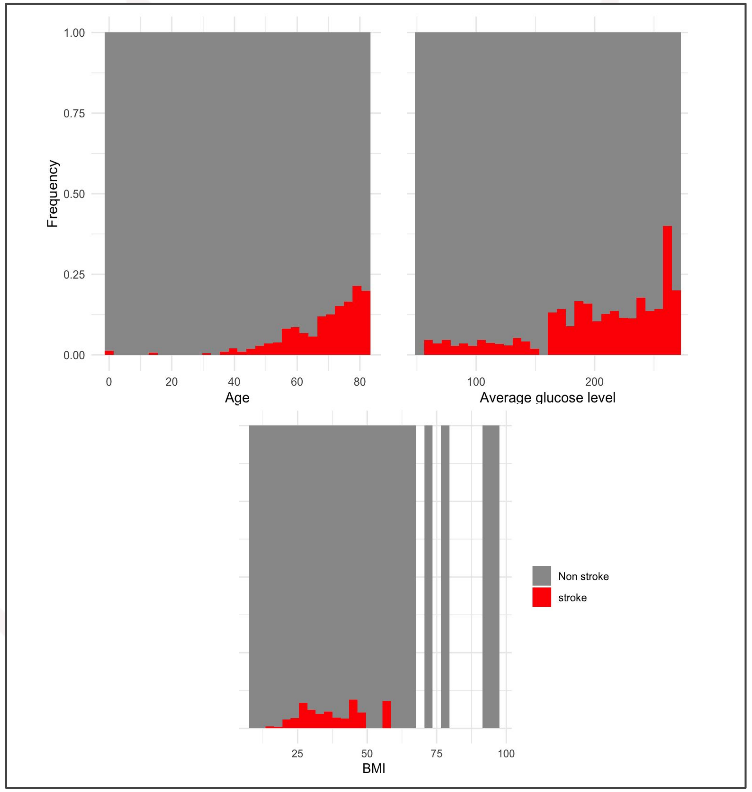
Exploratory Data Analysis (EDA)

Variable
Gender
Age
Hypertension
Heart Disease
Ever Married
Work Type
Residence Type
Avg Glucose Level
BMI
Smoking status
Stroke

Category variables



Numeric variables



Predictive Modeling

Logistic Regression Model

Model	#Variables	Threshold*	AUC	Accuracy	Sensitivity	Specificity	TP	TN	FP	FN	TPR	FPR
Logistic model	4	0.0388	0.8207	0.6954	0.8267	0.6886	62	1004	454	13	83%	31%
Lasso model	7	0.0523	0.8216	0.7260	0.8000	0.7222	60	1053	405	15	80%	28%
Group Lasso model	15	0.0403	0.8209	0.6967	0.8267	0.6900	62	1006	452	13	82%	31%

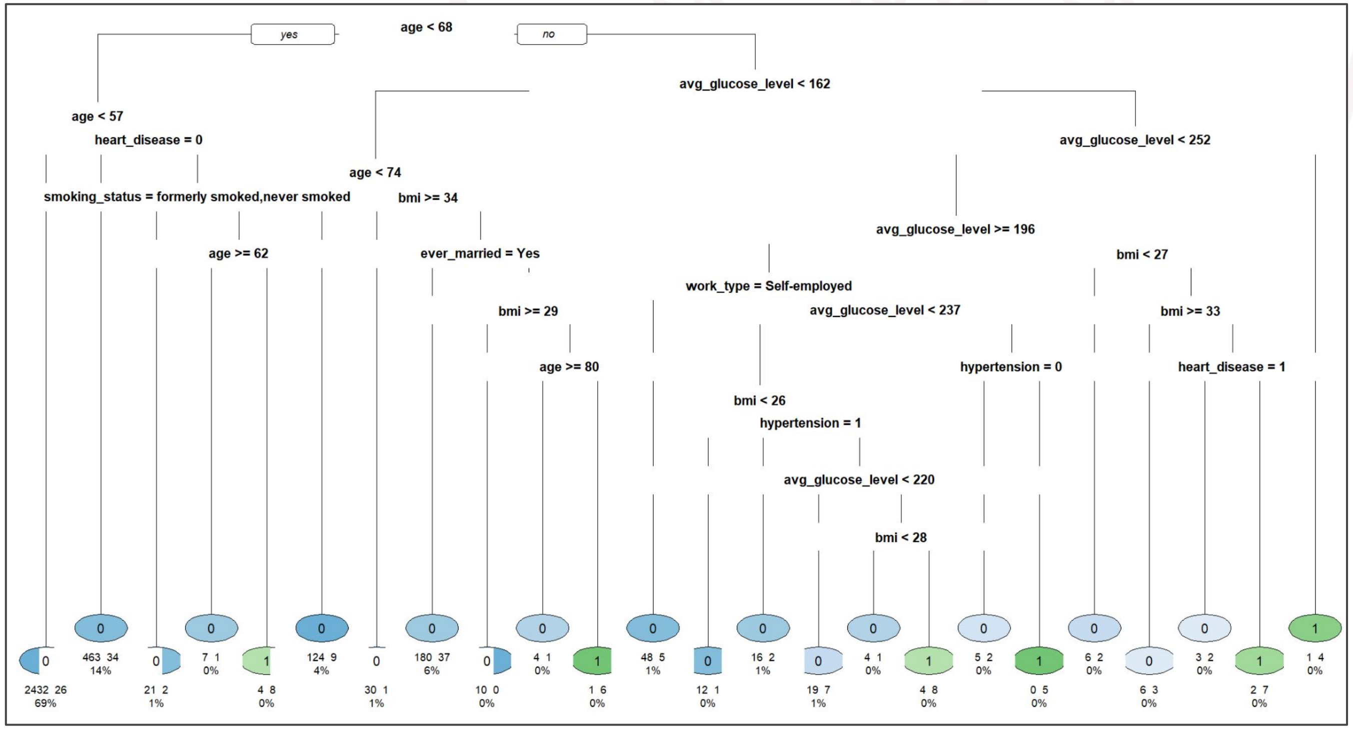
(*) Threshold that maximizes Youden's index (sensitivity - (1-specificity))

Classification And Regression Tree (CART) Model

Criteria	Parameters	Value
AUC	**cp**	**0.002**
	minsplit	**5**
	minbucket	5
	mindepth	13
	loss of false positive	1
	loss of false negative	1
Accuracy	Threshold	0.81

Model	AUC	Accuracy
CART model	0.7825	0.9511
CART model with pruning	0.7722	0.9511

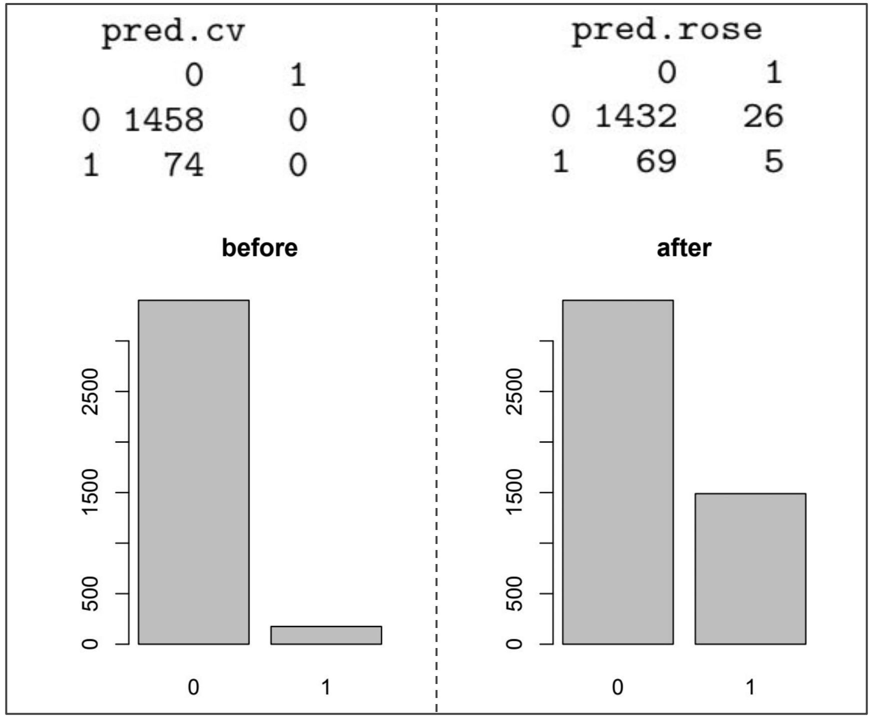
Pruned CART model at cp = 0.005



Predictive Modeling

Random Forest Model

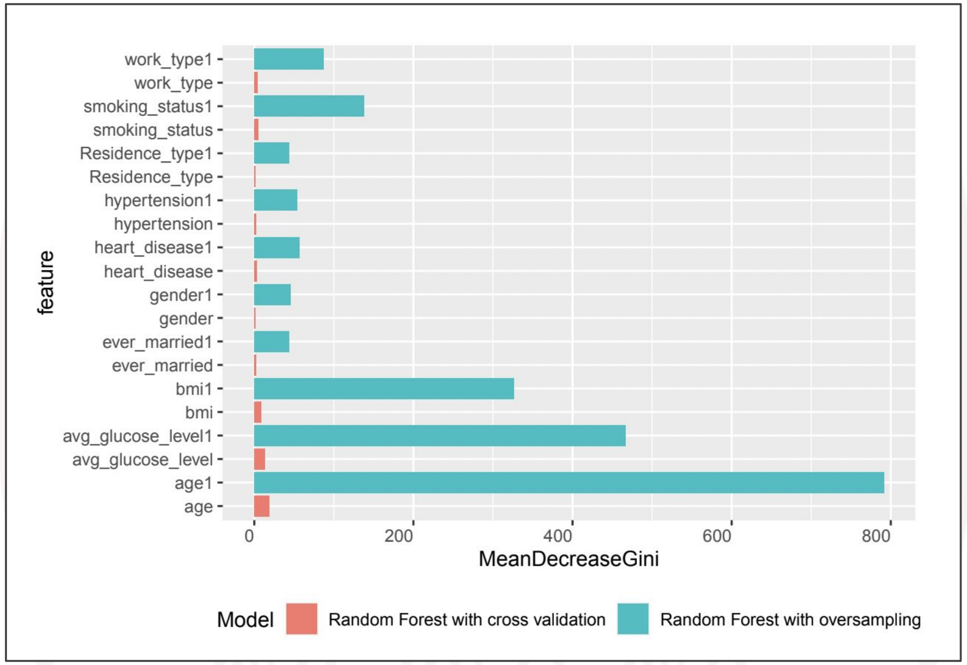
Before oversampling **After oversampling**

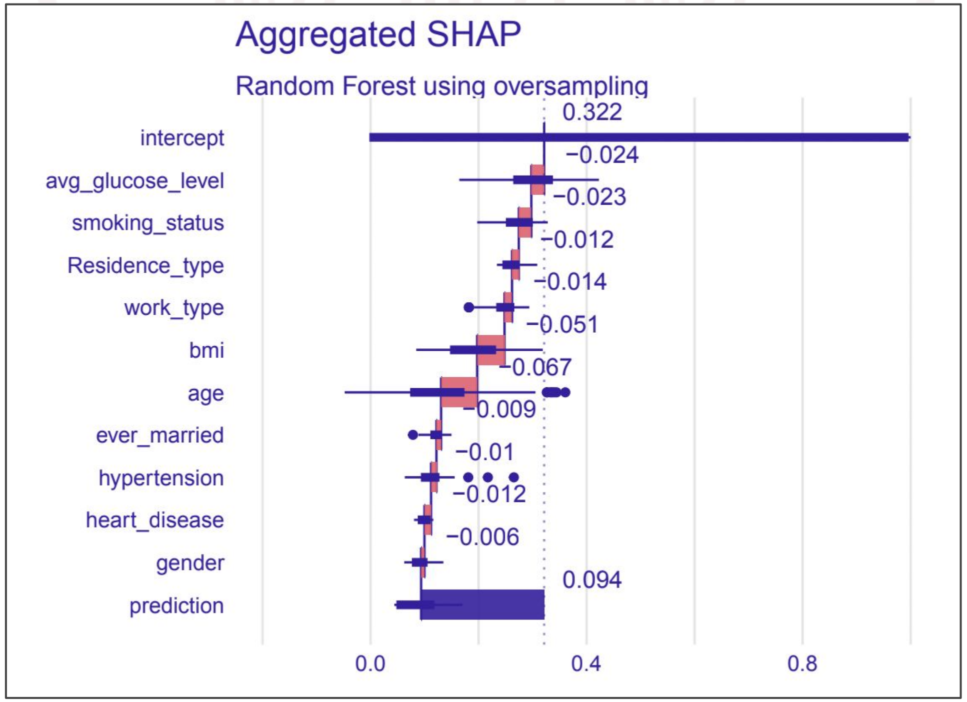


```
pred.cv
         0    1
0     1458    0
1       74    0
```

```
pred.rose
         0    1
0     1432   26
1       69    5
```

	Accuracy	AUC	TPR	FPR
Before	0.9517	0.7960	0	0
After	0.9380	0.8040	7%	2%





Conclusion of final model

- **Random forest** with oversampling and **Logistic regression** achieve the **highest performance**
- **CART** with pruning showed good interpretability but overall performance was comparatively lower
- **The logistic regression with group lasso** had a **high TPR of 82%** and a relatively **high FPR of 31%**

A high TPR can identify high-risk patients and facilitate early interventions

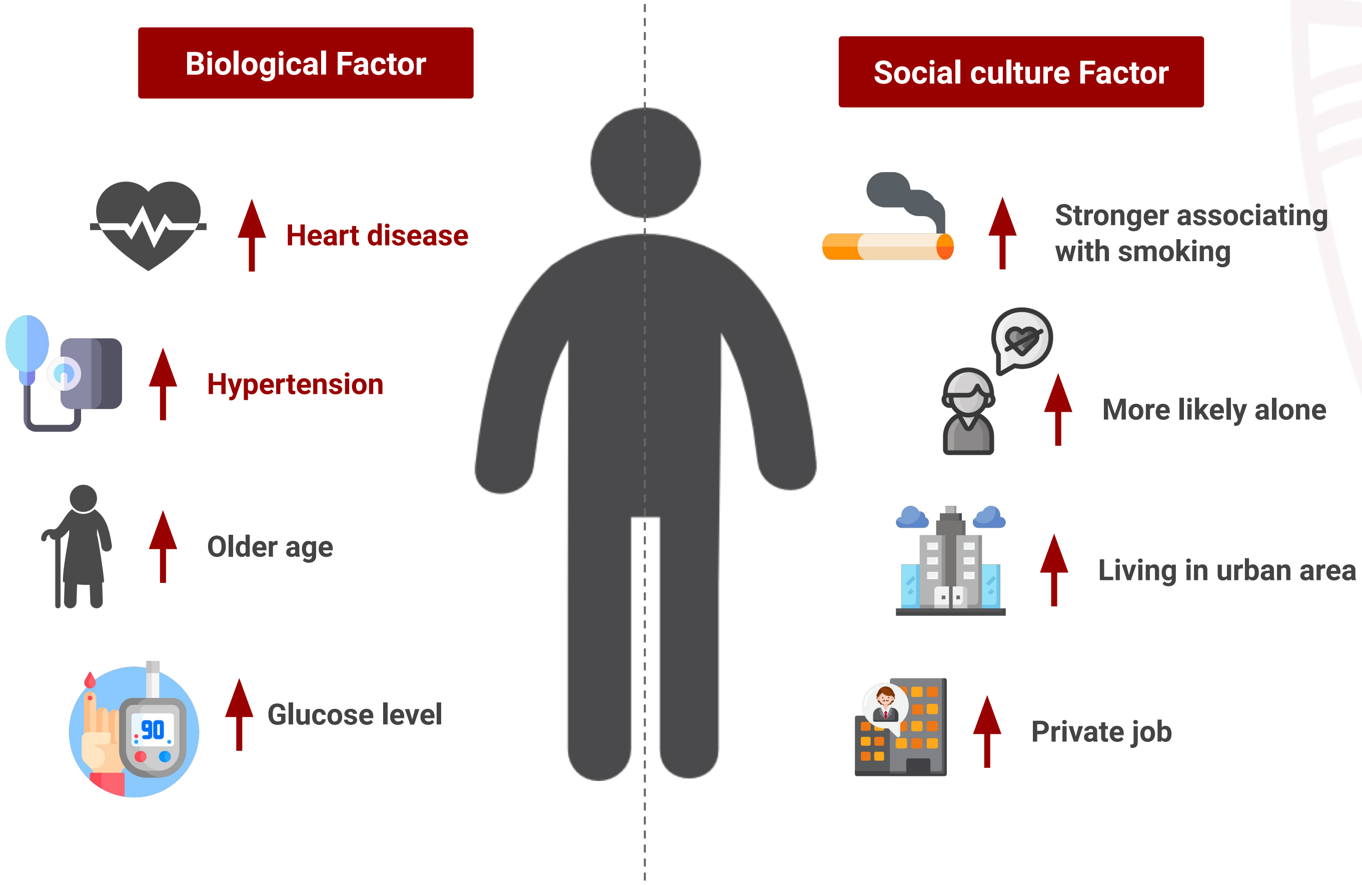
A high FPR can lead to unnecessary medical interventions, increased healthcare costs, and patient anxiety.

- The study **prioritized high TPR** was the most suitable choice for predicting stroke risk in patients.

Model	AUC	Accuracy	TP	TN	FP	FN	TPR	FPR
Logistic model	0.8207	0.6954	62	1004	454	13	83%	31%
Logistic model with Lasso	0.8216	0.7260	60	1053	405	15	80%	28%
Logistic model with Group Lasso	0.8209	0.6967	62	1006	452	13	82%	31%
CART model	0.7825	0.9511	3	1455	3	72	4%	0.2%
CART model with pruning	0.7722	0.9511	3	1455	3	72	4%	0.2%
Random forest	0.7960	0.9517	0	1458	0	74	0	0
Random Forest with Oversampling	0.8040	0.9380	5	1432	26	69	7%	2%

Implications and Recommendations

Risk Factors for experienced Stroke



Biological Factor

- ↑ Heart disease
- ↑ Hypertension
- ↑ Older age
- ↑ Glucose level

Social culture Factor

- ↑ Stronger associating with smoking
- ↑ More likely alone
- ↑ Living in urban area
- ↑ Private job

- **Develop targeted interventions**
encouraging regular check-ups and health screenings to monitor heart conditions, blood pressure, and glucose levels

- **Develop public awareness campaigns**
enhancing social connections, engaging in physical activity, and quitting smoking

- **Develop policies in urban areas**
promoting physical activity and outdoor recreation in urban areas and monitoring air quality

The prediction results should not be used as a substitute for thorough medical diagnostics.

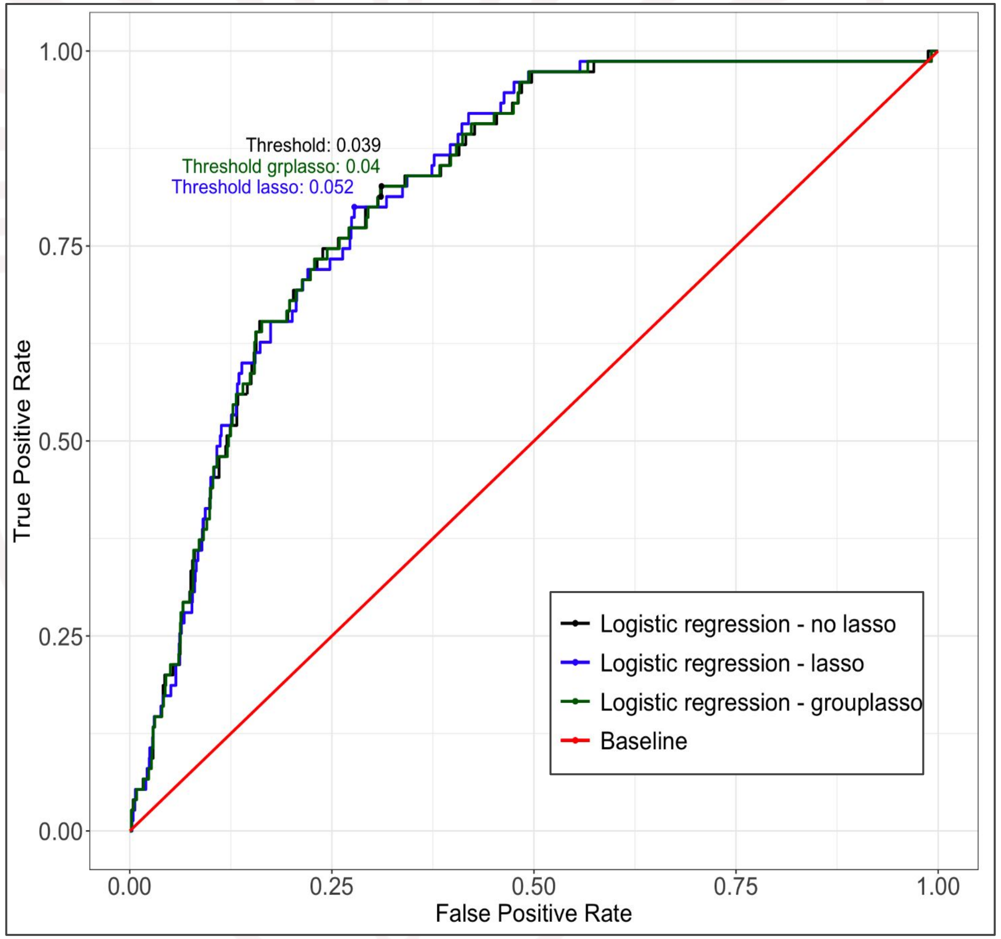
Thank you

Q&A

Appendix

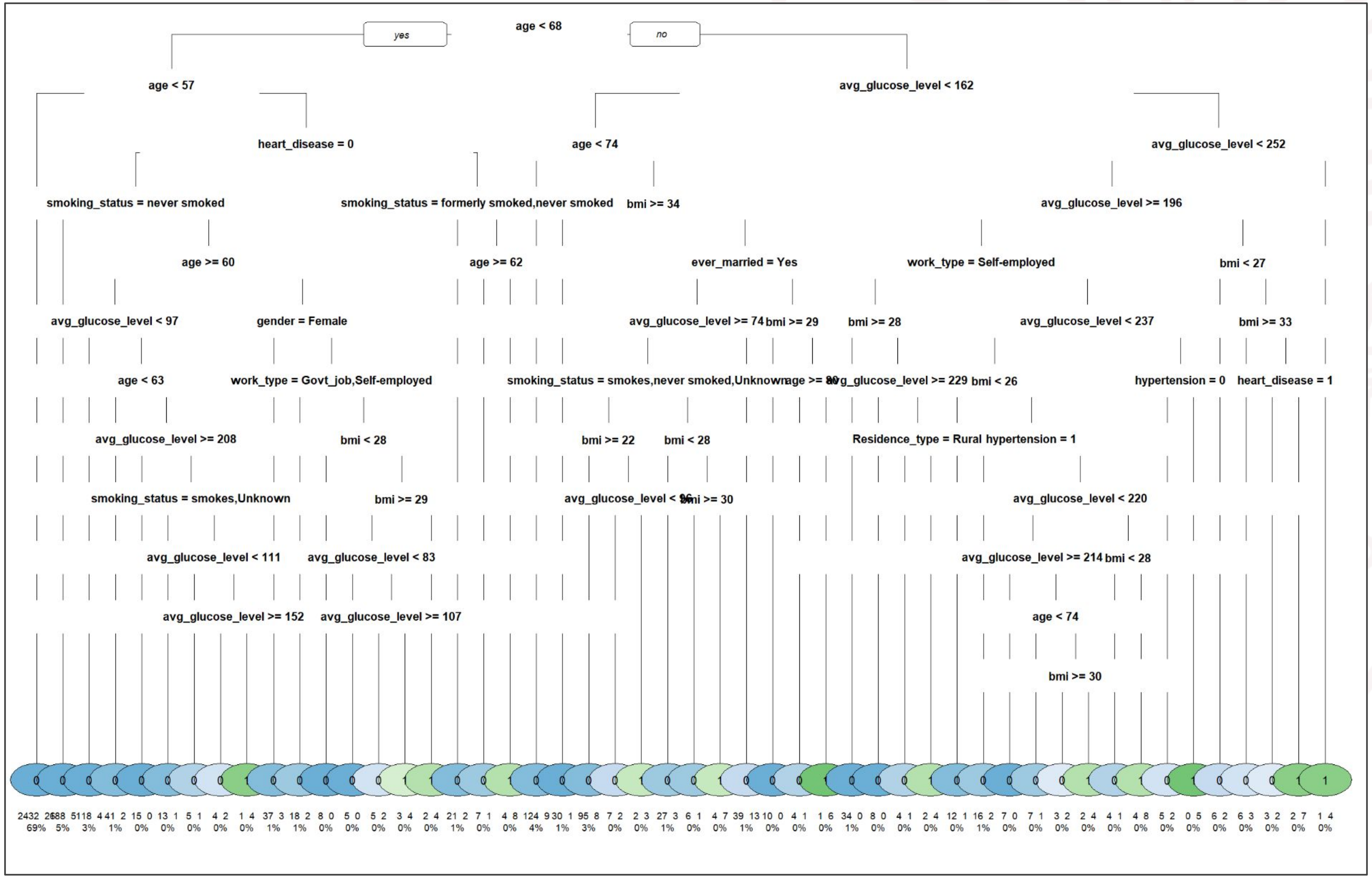
Logistic Regression Model

	Logistic regression		
	No Lasso	**With Lasso**	**With Group Lasso**
(Intercept)	-7.6419	-0.7138	-7.554
genderMale			0.1003
age	0.0753	0.0649	0.0735
hypertension1		0.1409	0.2109
heart_disease1	0.4478	0.4699	0.4533
ever_marriedNo			0.2154
work_typeGovt_job			-0.1318
work_typeSelf-employed	-0.4225	-0.1931	-0.3850
work_typeNever_worked			-0.4716
work_typechildren			0.0439
Residence_typeRural		-0.0494	-0.1526
avg_glucose_level	0.0055	0.0050	0.0054
bmi			-0.0020
smoking_statusformerly smoked			-0.0820
smoking_statusnever smoked		-0.2270	-0.3971
smoking_statusUnknown			-0.1020
Number of significant variables	**4**	**7**	**15**
AUC	**0.8206**	**0.8216**	**0.8209**



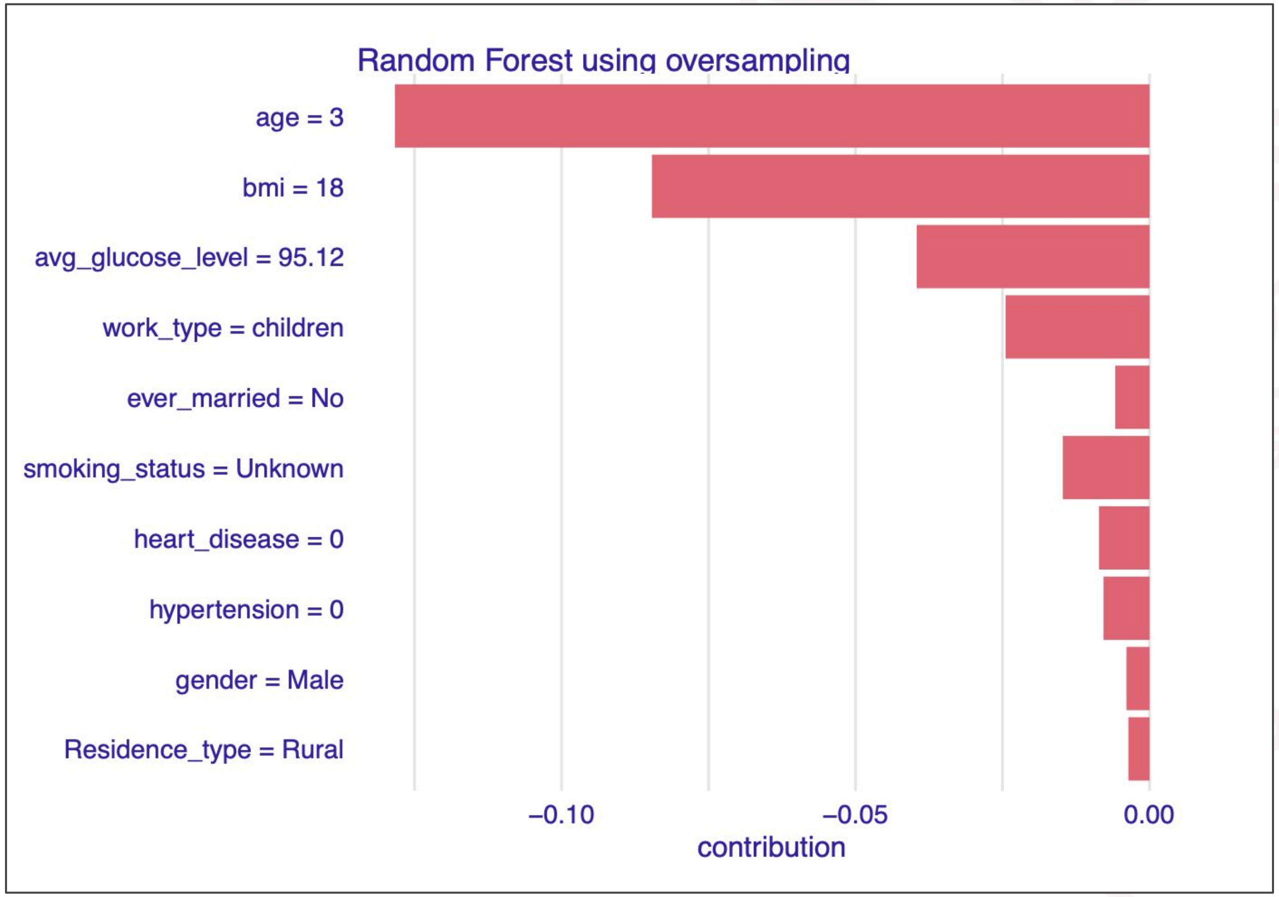
Classification And Regression Tree (CART) Model

Full CART Tree



Random Forest Model

Shapley values for individual observations


Random Forest using oversampling
age = 3
bmi = 18
avg_glucose_level = 95.12
work_type = children
ever_married = No
smoking_status = Unknown
heart_disease = 0
hypertension = 0
gender = Male
Residence_type = Rural
contribution