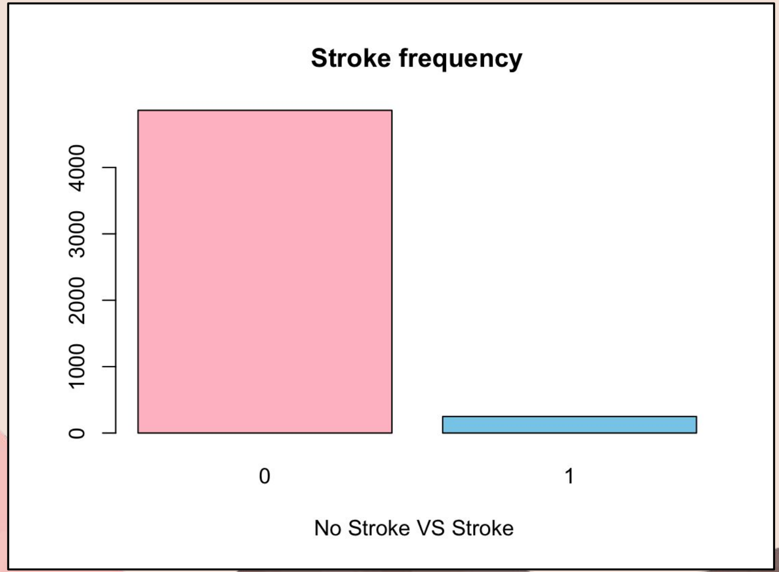
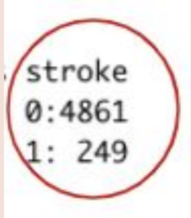
Stroke Prediction Analysis

Ngoc Nguyen

DATA INFORMATION

- 12 variables with 5110 observations





Highly **unbalanced** data distribution

May result in false accuracy at very high percentage of 0, which is healthy patients.

LOGISTIC REGRESSION

```
Call:
glm(formula = stroke ~ . - bmi, family = binomial, data = stroke_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2359  -0.3260  -0.1657  -0.0910   3.4697

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -6.545e+00  7.754e-01  -8.441  < 2e-16 ***
id                         4.368e-06  3.560e-06   1.227  0.21992
genderMale                -3.293e-02  1.552e-01  -0.212  0.83202
genderOther               -1.069e+01  1.455e+03  -0.007  0.99414
age                        7.613e-02  6.240e-03  12.200  < 2e-16 ***
hypertension1              4.519e-01  1.790e-01   2.525  0.01157 *
heart_disease1             2.802e-01  2.124e-01   1.320  0.18698
ever_marriedYes           -2.661e-01  2.478e-01  -1.074  0.28286
work_typeGovt_job         -1.147e+00  8.422e-01  -1.362  0.17317
work_typeNever_worked     -1.062e+01  4.018e+02  -0.026  0.97891
work_typePrivate          -1.065e+00  8.232e-01  -1.293  0.19587
work_typeSelf-employed    -1.403e+00  8.481e-01  -1.655  0.09794 .
Residence_typeUrban        4.857e-02  1.515e-01   0.320  0.74860
avg_glucose_level          3.539e-03  1.293e-03   2.736  0.00621 **
smoking_statusnever smoked -1.883e-01  1.938e-01  -0.972  0.33111
smoking_statussmokes       2.171e-01  2.295e-01   0.946  0.34420
smoking_statusUnknown     -1.747e-01  2.351e-01  -0.743  0.45761
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1655.9  on 4087  degrees of freedom
Residual deviance: 1307.2  on 4071  degrees of freedom
AIC: 1341.2

Number of Fisher Scoring iterations: 14
```

$$\text{logit}(y) = -6.545 + 0.000004368X_1 - 0.03293X_2 + \ldots + 0.2171X_{15} - 0.1747X_{16}$$

- Run variable selection

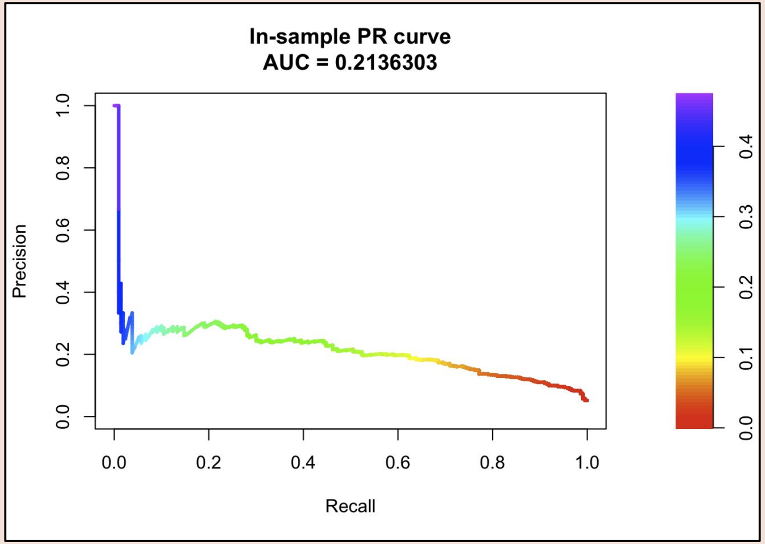
⇒ Find the best model is **backward selection** model

$$\text{new logit}(y) = -7.382350 + 0.069275X_1 + 0.457505X_2 + 0.341837X_3 + 0.003440X_4$$

IN SAMPLE PERFORMANCE

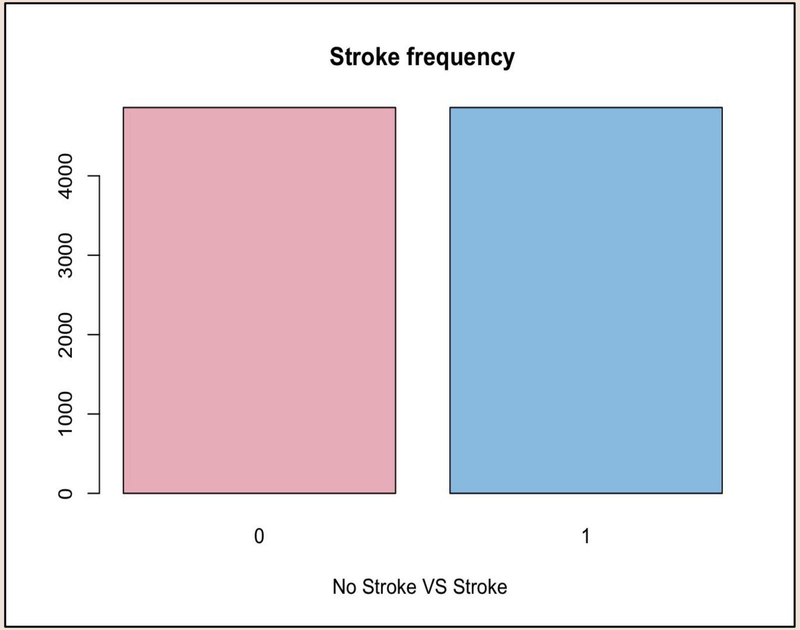
Cut-off probability: 0.2	**Predicted**	
Truth	0	1
0	3704	174
1	**150**	**60**

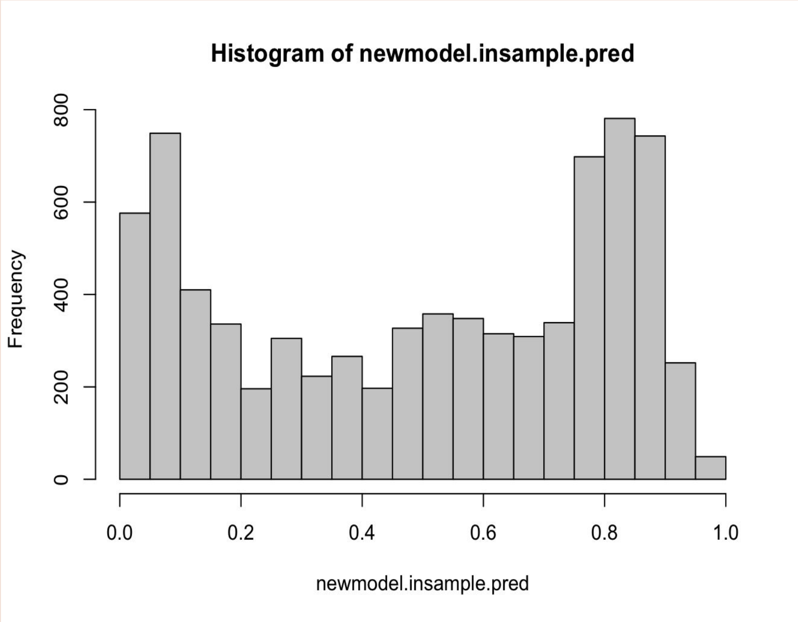
150 out of 210 stroke patients will pass the stroke detection
⇒ **Not very good model**



RESAMPLING DATA

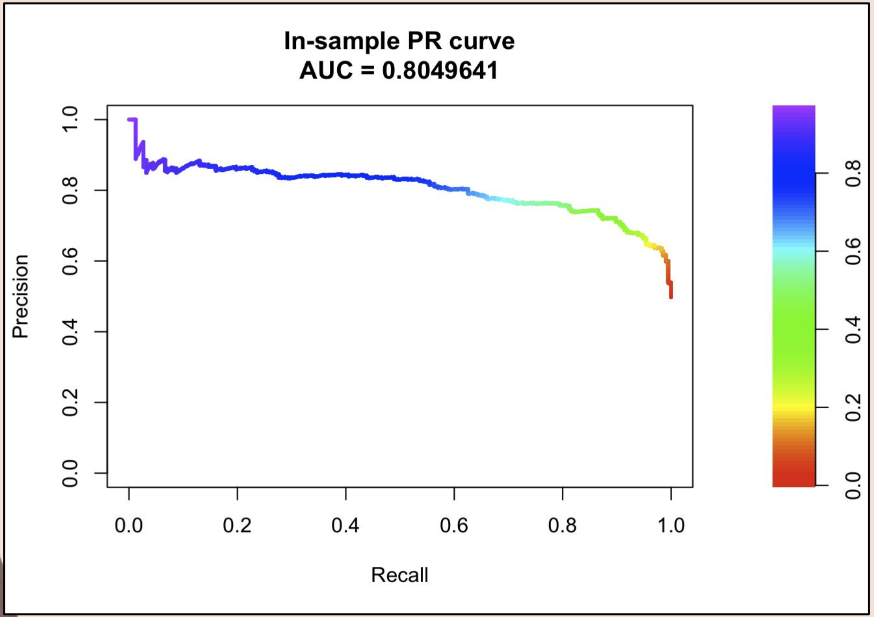
- Oversampling data using ROSE method





⇒ The data is now **balanced** with 4861 observations in each value

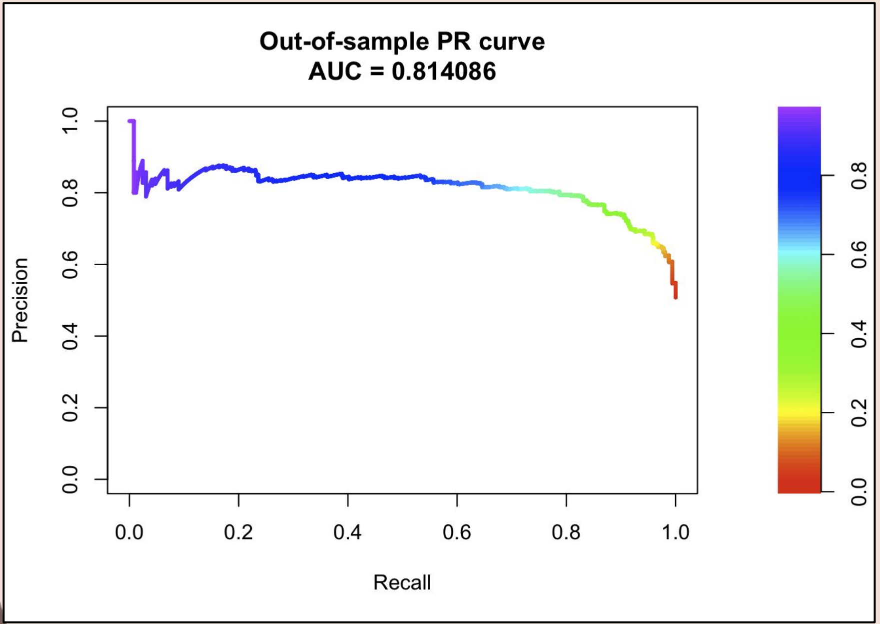
IN SAMPLE PERFORMANCE WITH BALANCED DATA



Cut-off probability: 0.2	**Predicted**	
Truth	0	1
0	1892	**2012**
1	**179**	3694

- AUC rises from **0.214** to **0.805**
- fewer false negatives - the model catches more stroke, unhealthy patients

OUT-OF-SAMPLE PERFORMANCE



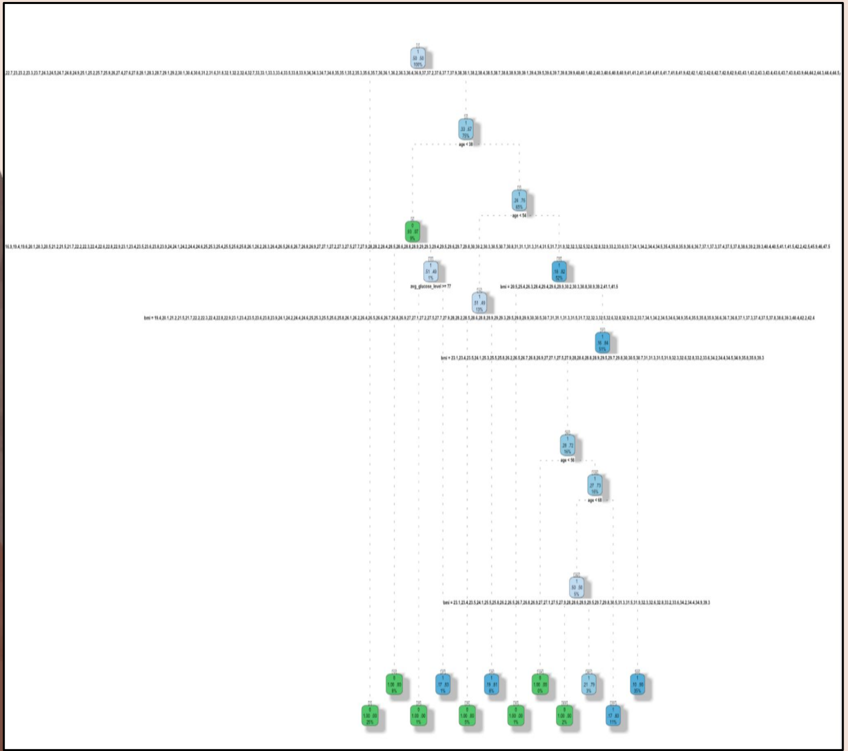
Cut-off probability: 0.2	**Predicted**	
Truth	0	1
0	472	**485**
1	**41**	947

947 out of **988** stroke, unhealthy patients are predicted correctly

73% of accuracy

DECISION TREE

Classification tree with **asymmetric cost**



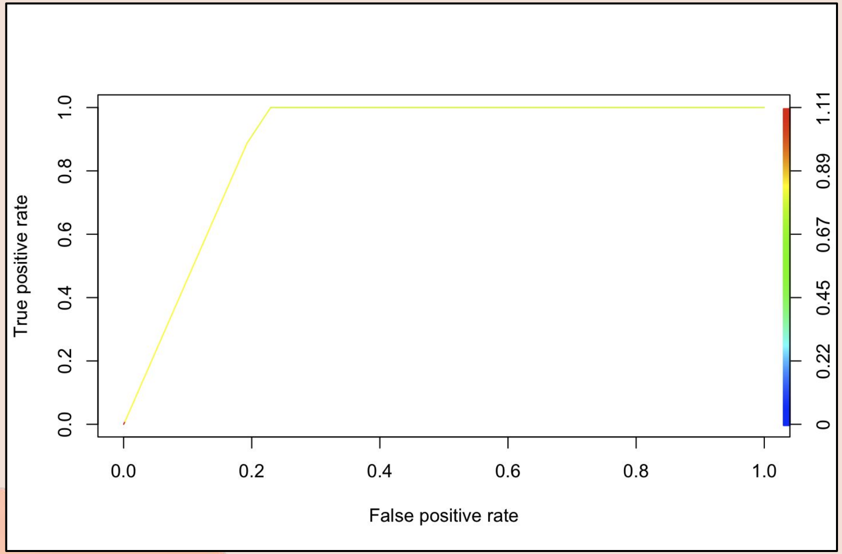
Selected variables: "**bmi**", "**avg_glucose_level**", and "**age**"

Symmetric cost	**Predicted**	
Truth	0	1
0	3201	703
1	33	3840

Asymmetric cost	**Predicted**	
Truth	0	1
0	3170	734
1	0	3873

DECISION TREE: OUT-OF-SAMPLE PERFORMANCE

Asymmetric cost	**Predicted**	
Truth	0	1
0	735	219
1	**0**	991



AUC = 0.8905
88.7% of accuracy

Better model than the Logistic Regression

RANDOM FOREST

- **In-sample performance with 80% training data**

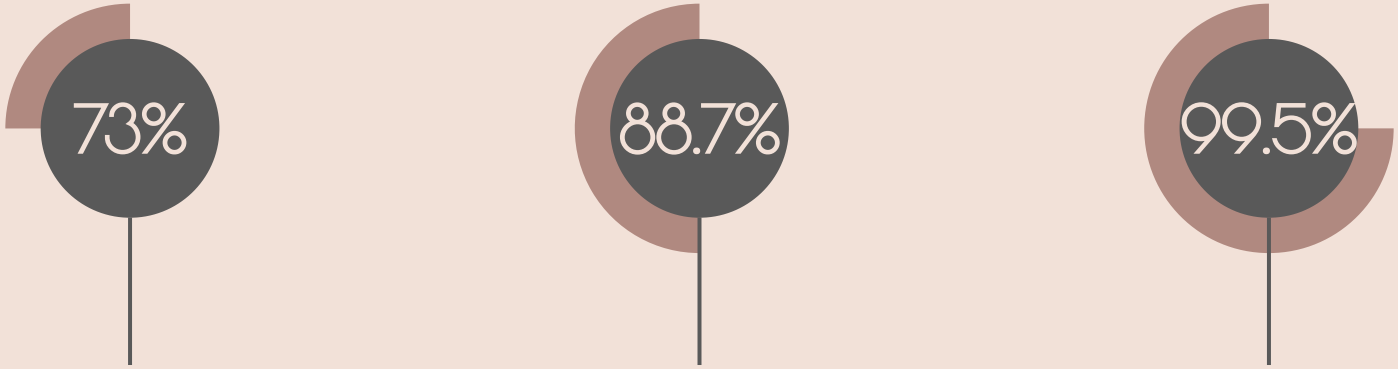
```
Call:
 randomForest(formula = stroke ~ ., data = stroke_train1, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.6%
Confusion matrix:
     0    1 class.error
0 3860   47  0.01202969
1    0 3870  0.00000000
```

- **Out-of-sample performance with 20% testing data**

	Predicted	
Truth	0	1
0	944	**10**
1	**0**	991

99.5% of accuracy

KEY FINDINGS



Logistic Regression | Decision Tree | Random Forest

- **Random forest** performs the best prediction model
- **Bmi, glucose level, and age** are the most influential stroke risk factors
- **Genders, work types, residence types** are not associated with stroke experience
- Balancing the data by applying oversampling or undersampling can improve the accuracy/**avoid sending potential stroke patients home**.

THANK YOU FOR LISTENING

Do you have any questions?