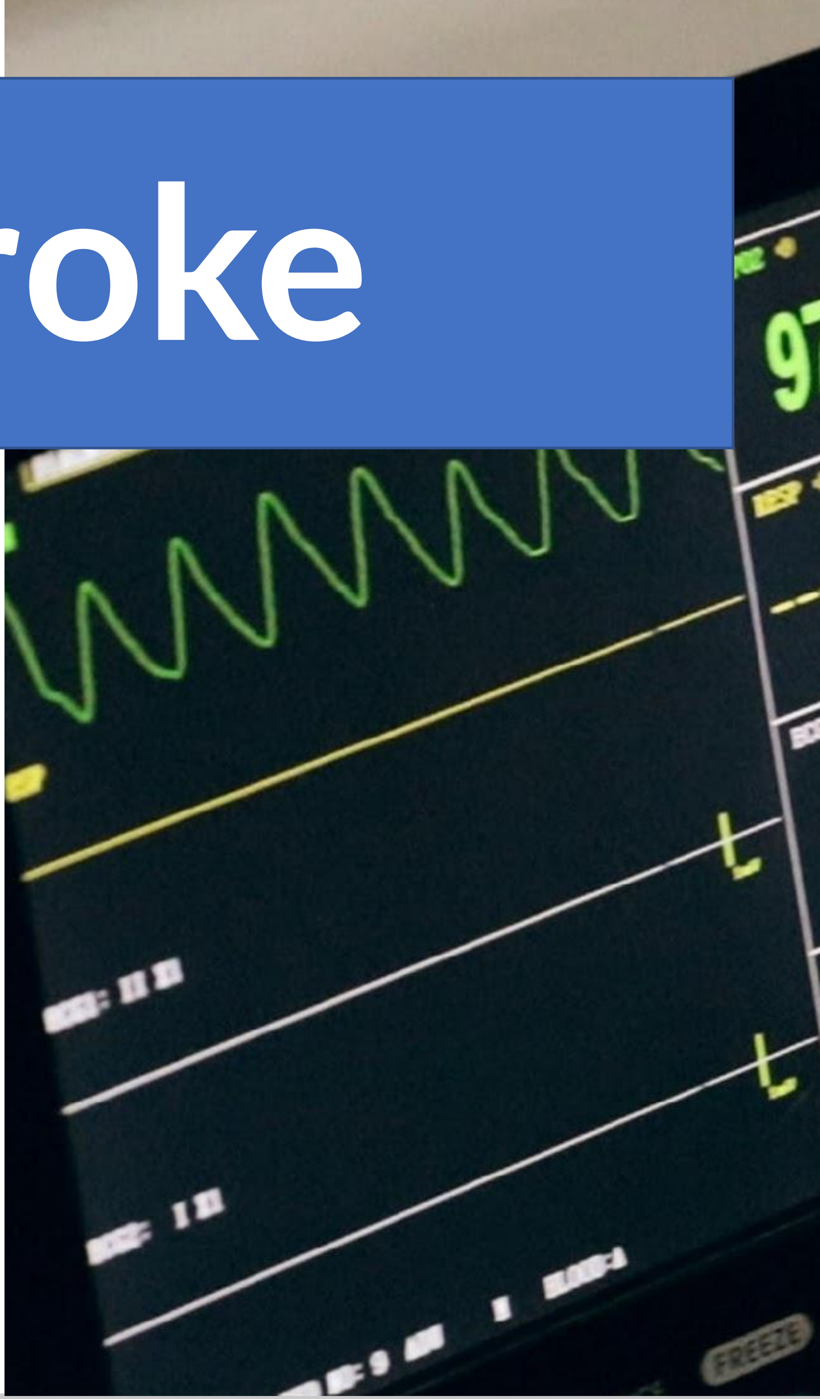
Predicting Stroke

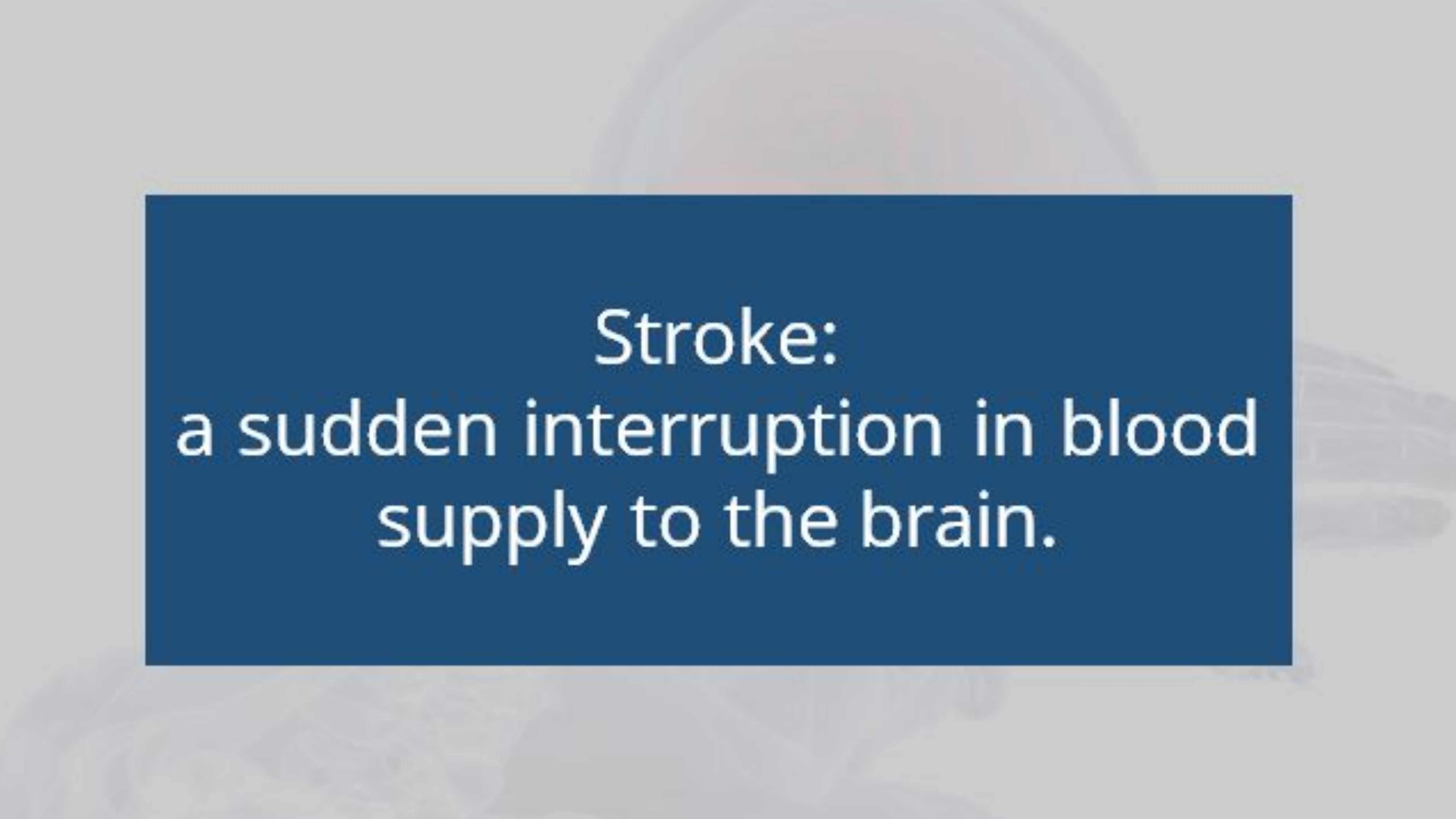


Predictive Modeling
Summer 2019

Ananya
Andrew
Josh
Karen
Reid
Vinay

Agenda

- Our problem / Review of dataset
- Exploratory Analysis
- Modeling
  - KNN
  - Logistic Regression
  - Tree Models
  - Boosting
  - Random Forest
  - Bagging
- Model Comparisons

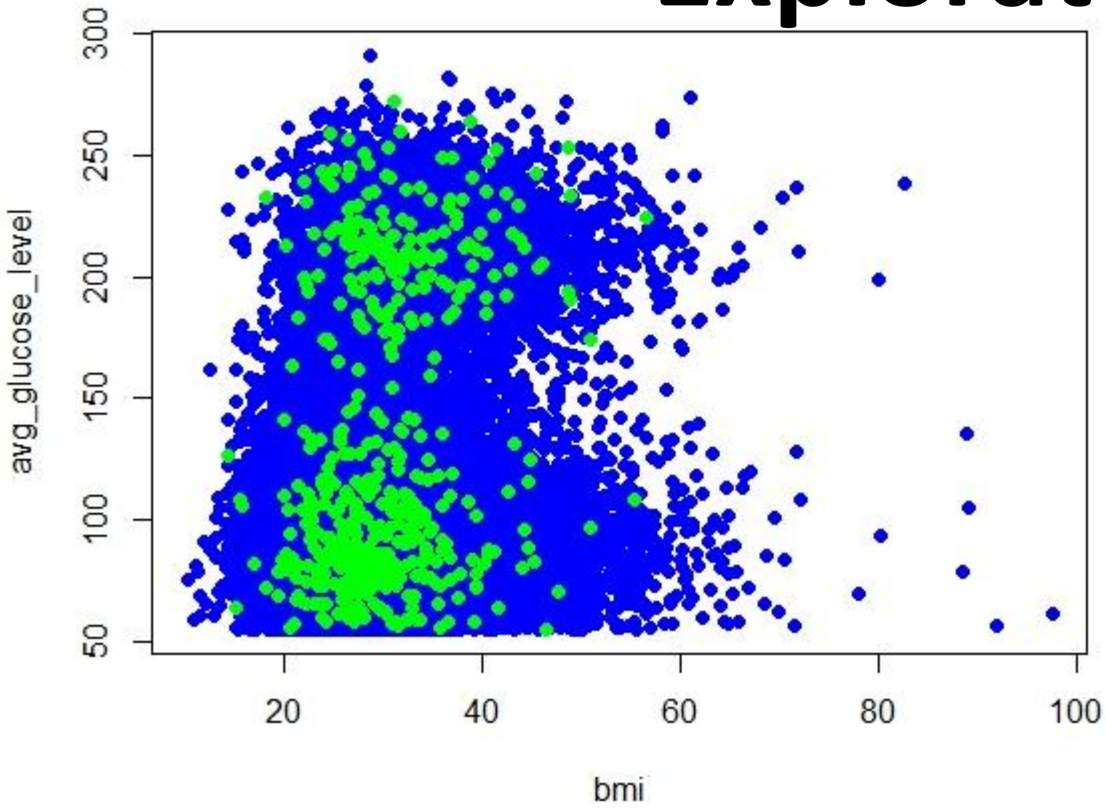


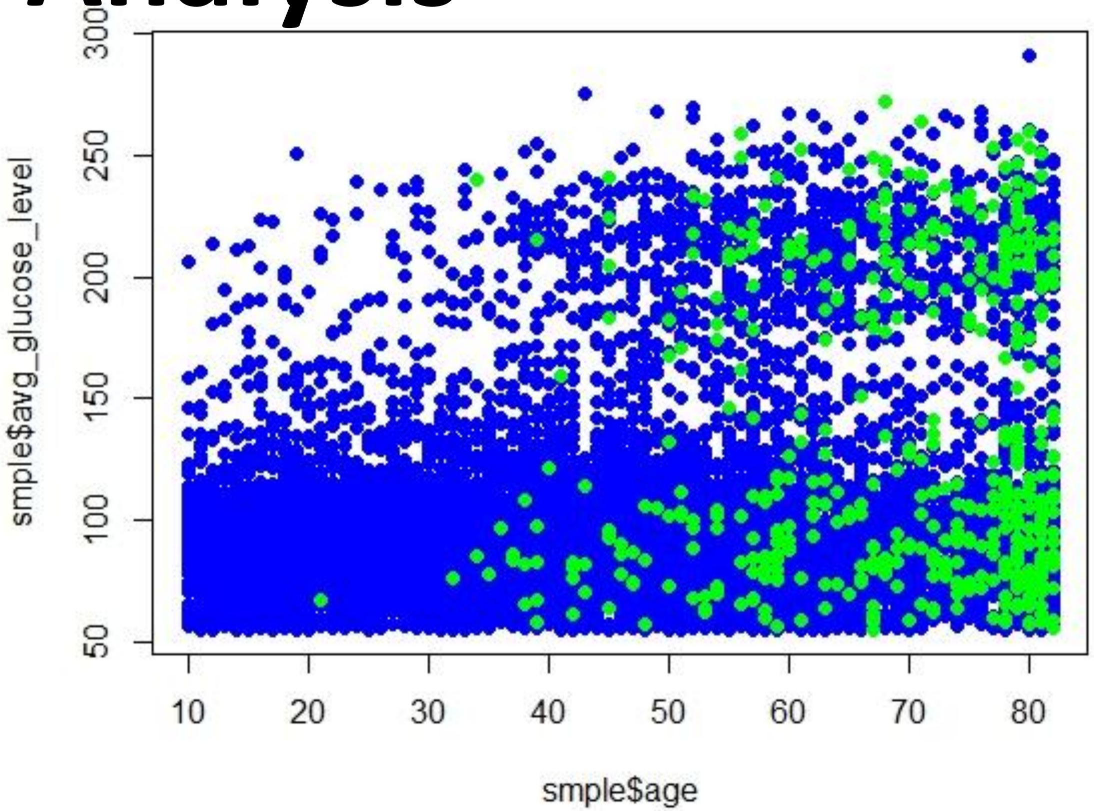
Stroke:
a sudden interruption in blood supply to the brain.

Introduction to our data

- 43,400 data points, representing patient data
- 12 variables
    - gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose levels, body mass index, smoking status, and whether or not they are stroke victims
- Data cleansing:
    - removed the 15 patients with missing gender data
    - removed patients under the age of 15

Exploratory Analysis

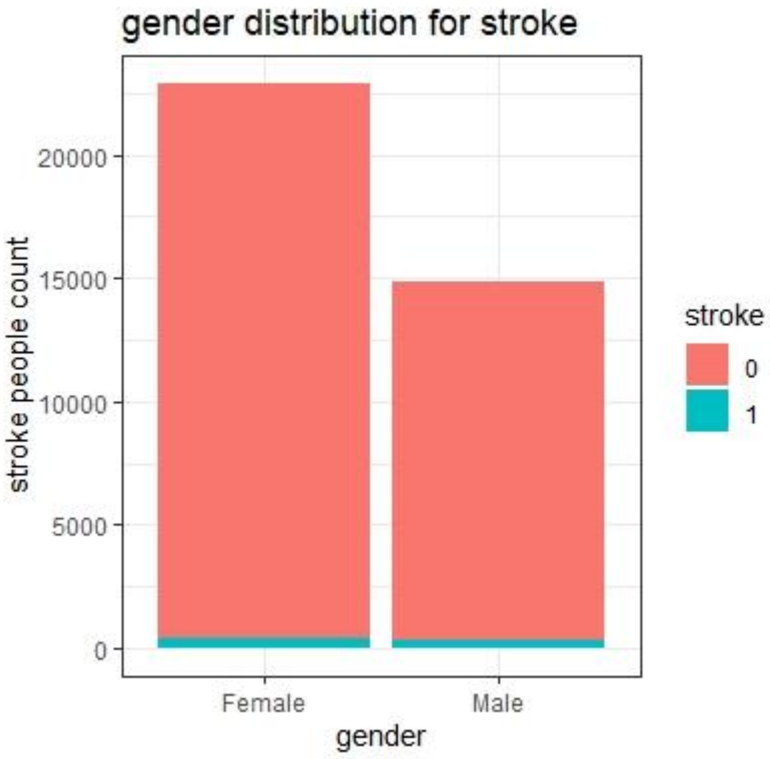




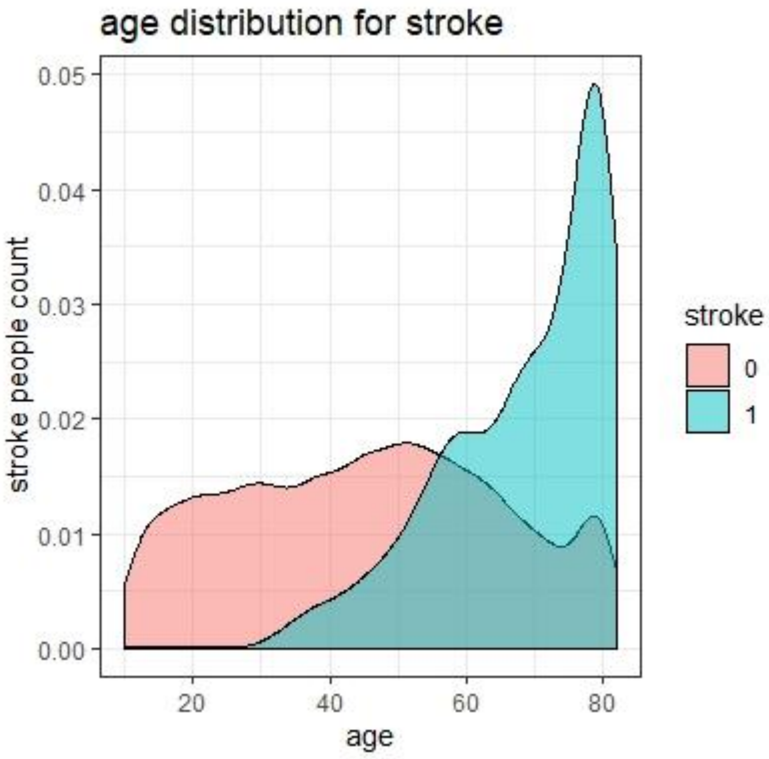
Strokes victims highlighted in green

- We were hoping to see if the stroke victims were clumped together under certain statistics

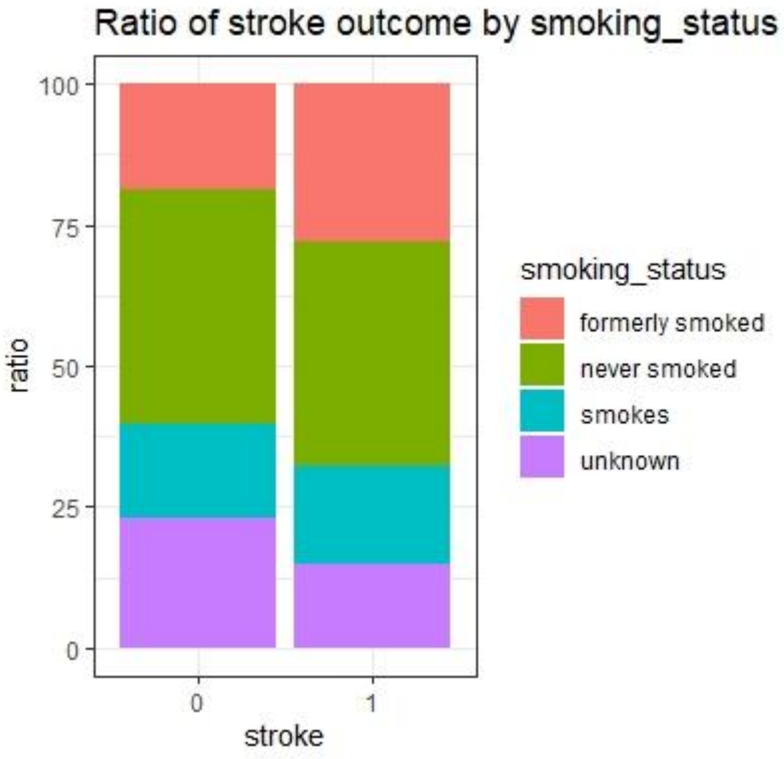
Exploratory Analysis



gender distribution for stroke



age distribution for stroke



Ratio of stroke outcome by smoking_status

Using Accuracy as a performance metric

Classification problem: varying penalties

Skewed Data: accuracy already at 98%

Performance metrics:

Precision: fraction of relevant instances among the retrieved instances
Recall: fraction of the relevant instances successfully retrieved
Misclassification Rate: fraction of incorrectly labeled instances

KNN

Three models tried:

1. All features
2. Only age and gender
3. All medical features
4. All lifestyle features + age + gender

Validated by 10 - fold Cross-Validation

Best value of K for KNN: 300

Threshold set to 0.03

Confusion Matrix

Predicted Actual	Non-Stroke	Stroke
Non-Stroke	3197	503
Stroke	12	60

Misclassification Rate: 0.112

Precision: 0.112

Recall: 0.883

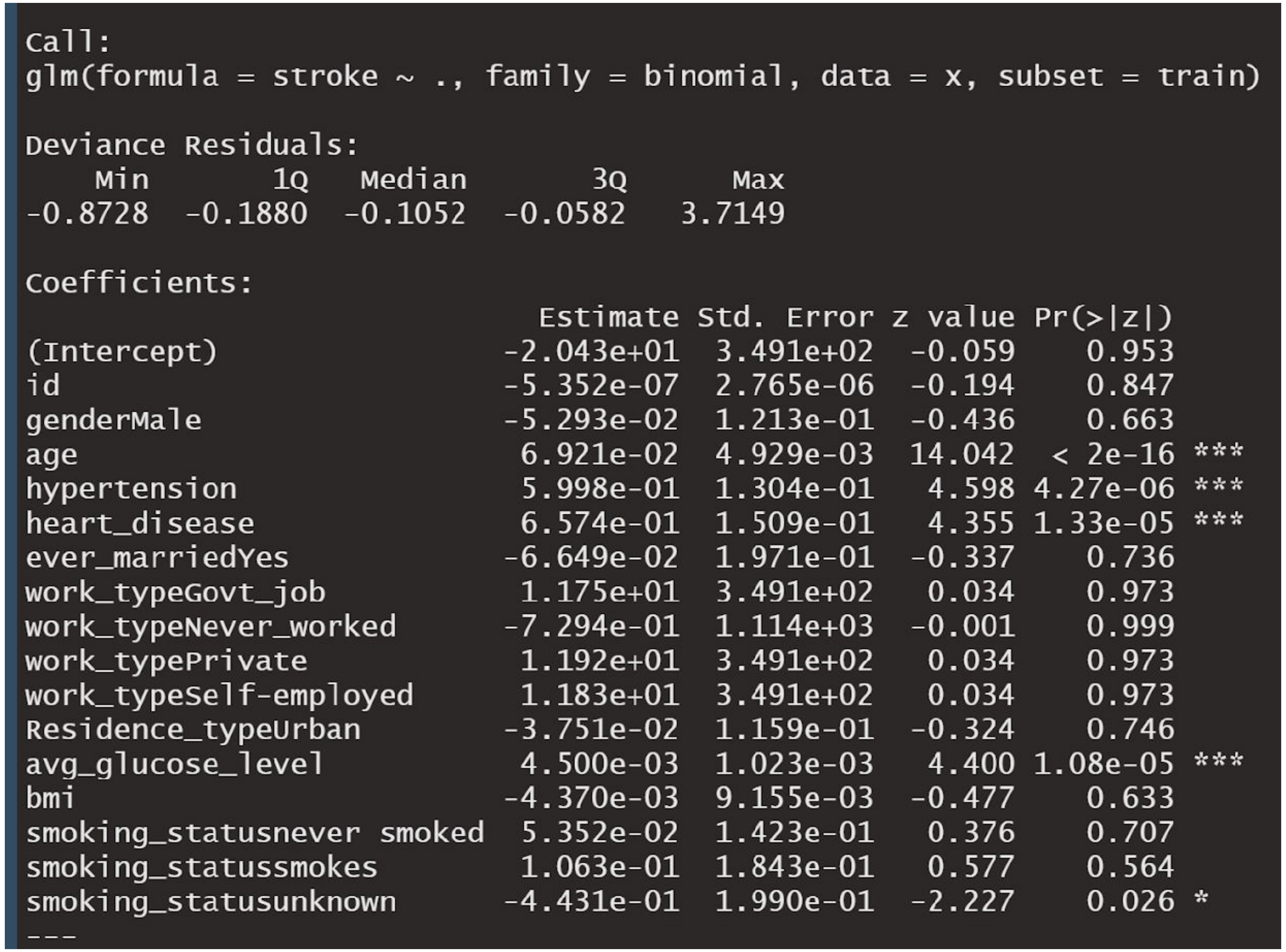
Logistic Regression Model

Imbalanced Data
1.7%

Threshold set at 0.03

Significant Predictors

Age
Hypertension
Heart Disease
Avg glucose level



```
Call:
glm(formula = stroke ~ ., family = binomial, data = x, subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8728  -0.1880  -0.1052  -0.0582   3.7149

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.043e+01  3.491e+02  -0.059    0.953
id                         -5.352e-07  2.765e-06  -0.194    0.847
genderMale                 -5.293e-02  1.213e-01  -0.436    0.663
age                         6.921e-02  4.929e-03  14.042  < 2e-16 ***
hypertension                5.998e-01  1.304e-01   4.598 4.27e-06 ***
heart_disease               6.574e-01  1.509e-01   4.355 1.33e-05 ***
ever_marriedYes            -6.649e-02  1.971e-01  -0.337    0.736
work_typeGovt_job           1.175e+01  3.491e+02   0.034    0.973
work_typeNever_worked      -7.294e-01  1.114e+03  -0.001    0.999
work_typePrivate            1.192e+01  3.491e+02   0.034    0.973
work_typeSelf-employed      1.183e+01  3.491e+02   0.034    0.973
Residence_typeUrban        -3.751e-02  1.159e-01  -0.324    0.746
avg_glucose_level           4.500e-03  1.023e-03   4.400 1.08e-05 ***
bmi                        -4.370e-03  9.155e-03  -0.477    0.633
smoking_statusnever smoked  5.352e-02  1.423e-01   0.376    0.707
smoking_statussmokes        1.063e-01  1.843e-01   0.577    0.564
smoking_statusunknown      -4.431e-01  1.990e-01  -2.227    0.026 *
---
```

Logistic Regression Model

Model Evaluated on test data

Confusion Matrix

Predicted Actual	Non-Stroke	Stroke
Non-Stroke	15478	2985
Stroke	117	208

Misclassification Rate: 0.165

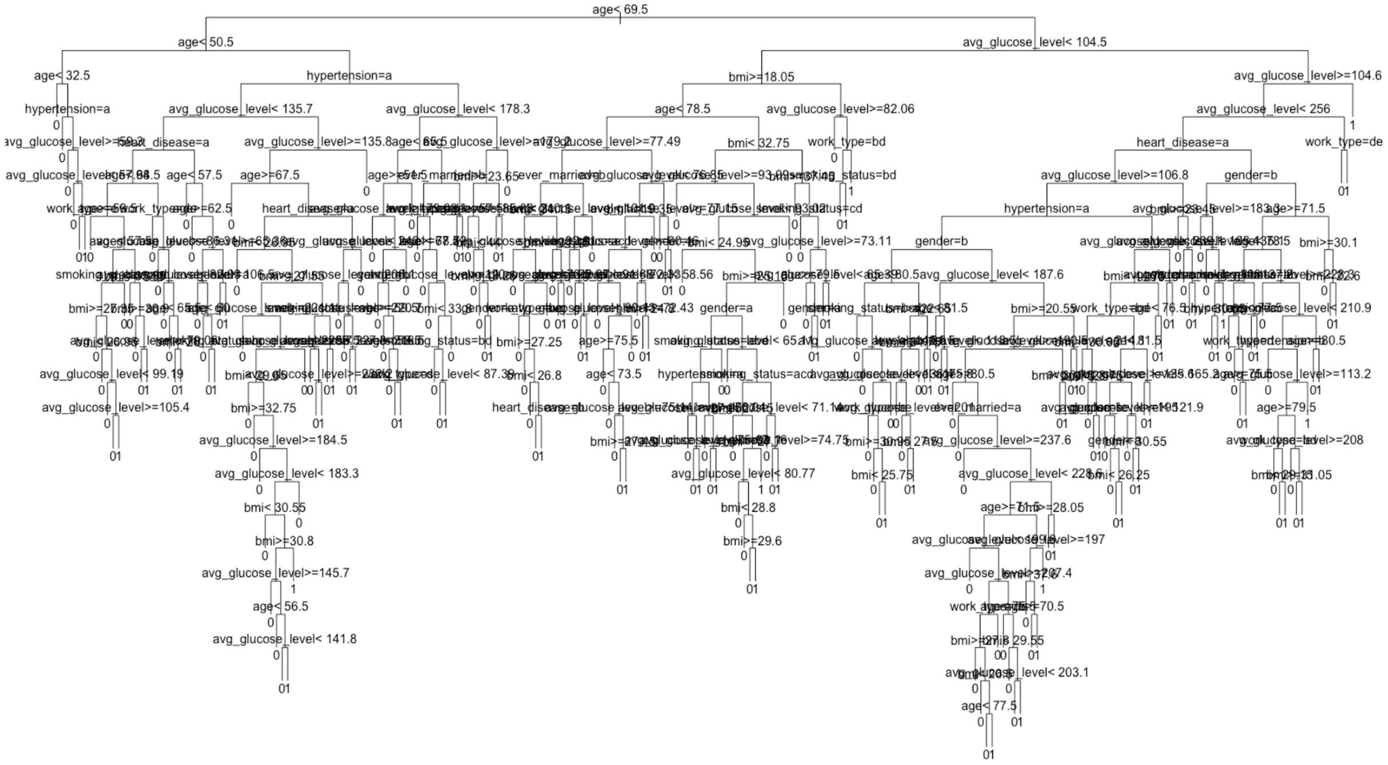
Precision: 0.069

Recall: 0.64

Classification - Tree Model

Resample: Unbalanced

- **Used resampling method to select train, validation and test data**
- **Train accounts for a half of data and validation and test accounts for a quarter of data.**



Confusion Matrix

Predicted / Actual	Non-Stroke	Stroke
Non-Stroke	9180	85
Stroke	158	6

- Misclassification Rate: 0.036
- Accuracy Rate: 0.974
- Recall Rate: 0.034
- Precision Rate: 0.06

Classification - Tree Model

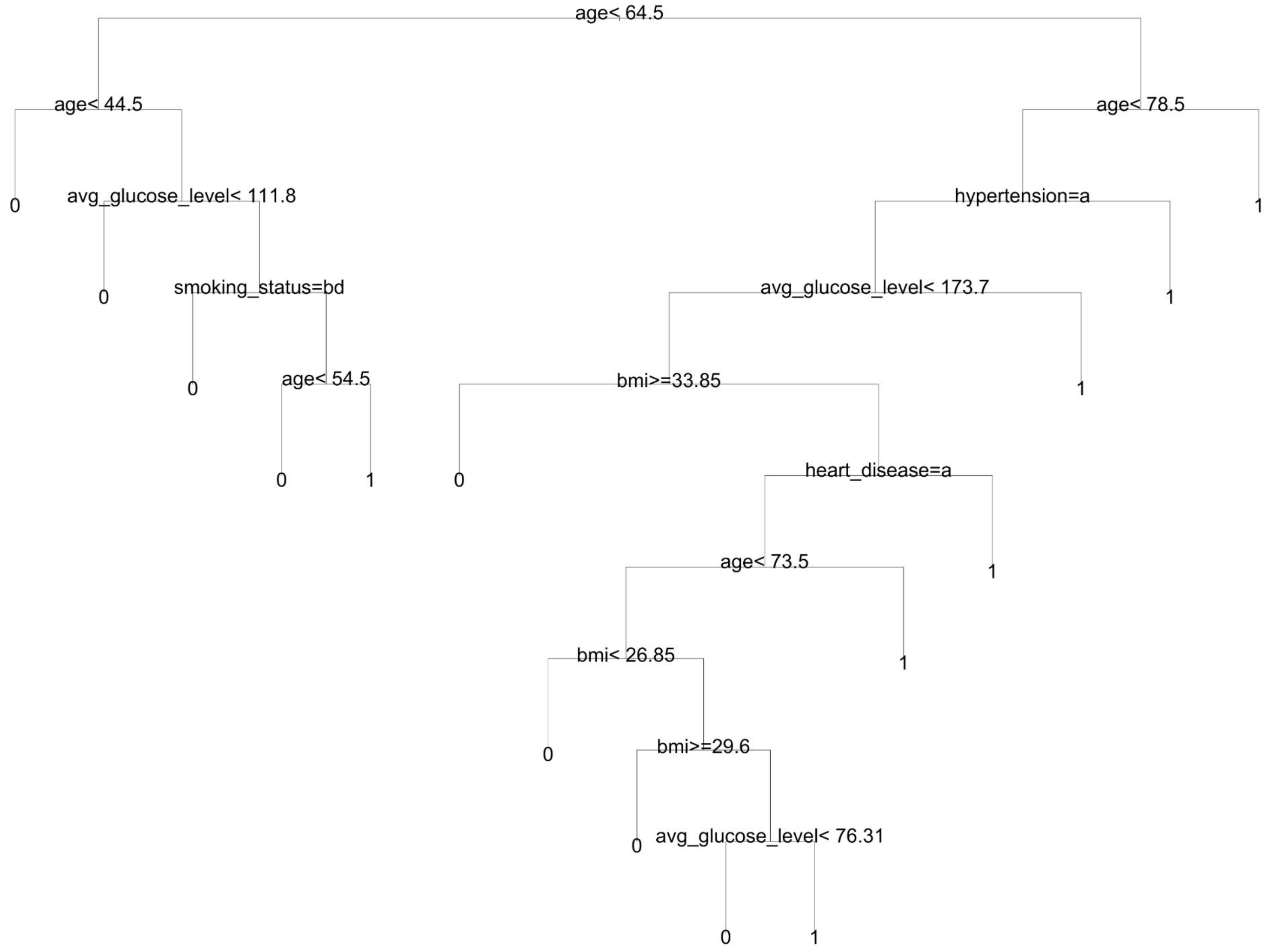
Construct balanced dataset

- Selected 500 stroke records and 1000 non-stroke records as the train dataset
- Divided the rest of data equally into validation and test

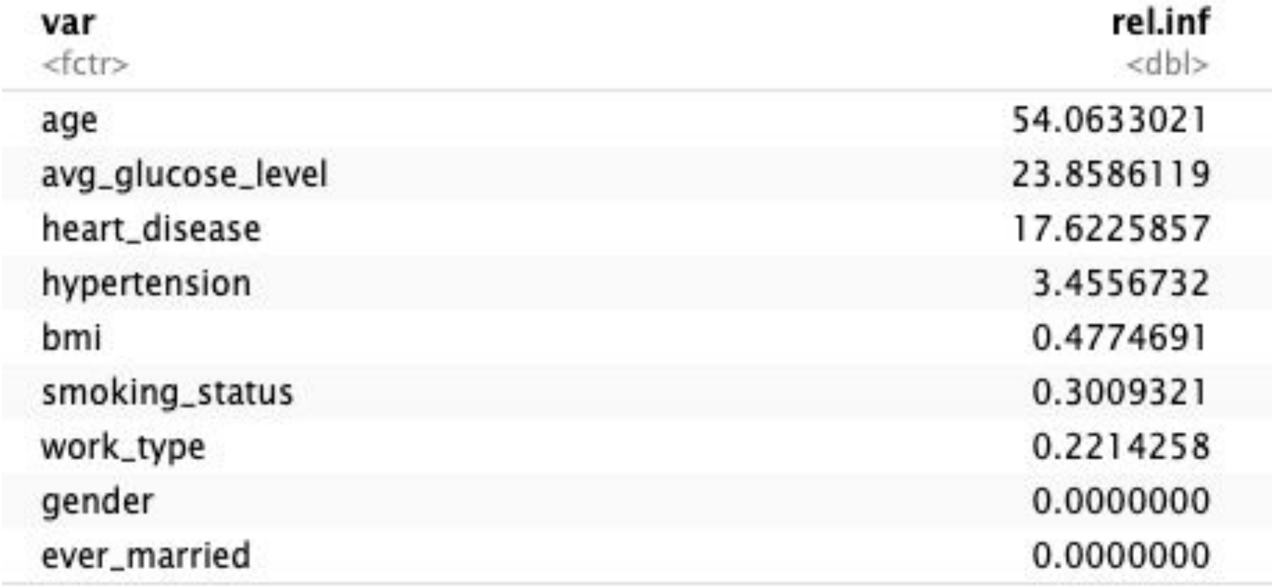
Confusion Matrix

Predicted Actual	Non-Stroke	Stroke
Non-Stroke	14443	3591
Stroke	25	50

- Misclassification Rate: 0.20
- Accuracy Rate: 0.80
- Recall Rate: 0.67
- Precision Rate: 0.014



Boosting



var <fctr>	rel.inf <dbl>
age	54.0633021
avg_glucose_level	23.8586119
heart_disease	17.6225857
hypertension	3.4556732
bmi	0.4774691
smoking_status	0.3009321
work_type	0.2214258
gender	0.0000000
ever_married	0.0000000

Confusion Matrix

Predicted Actual	Non-Stroke	Stroke
Non-Stroke	16319	2223
Stroke	162	155

Misclassification Rate: 0.13

Precision: 0.065

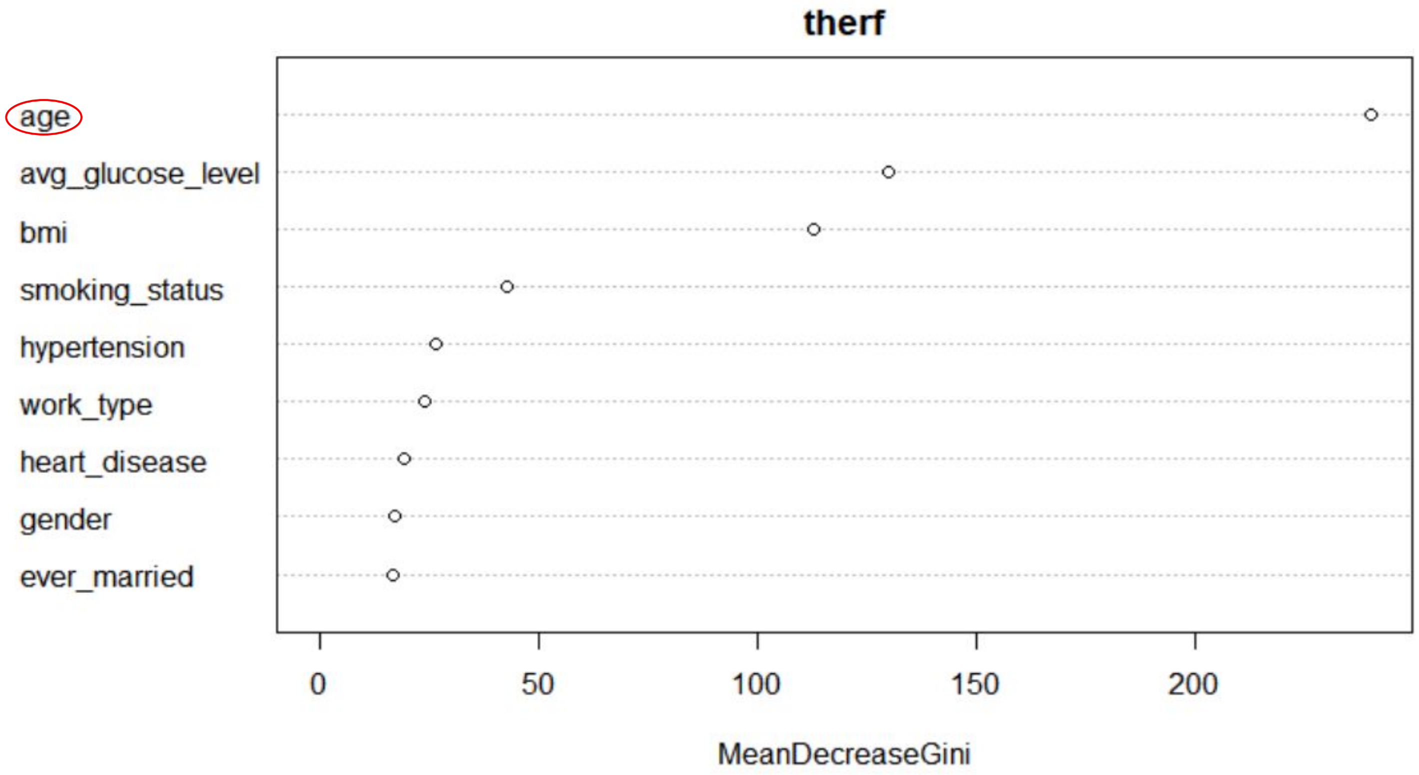
Recall: 0.490

Random Forest

Confusion Matrix

Actual	Non-Stroke	Stroke
Non-Stroke	15804	2956
Stroke	31	38

- Misclassification Rate=0.15
- Precision= 0.012
- Recall= 0.55

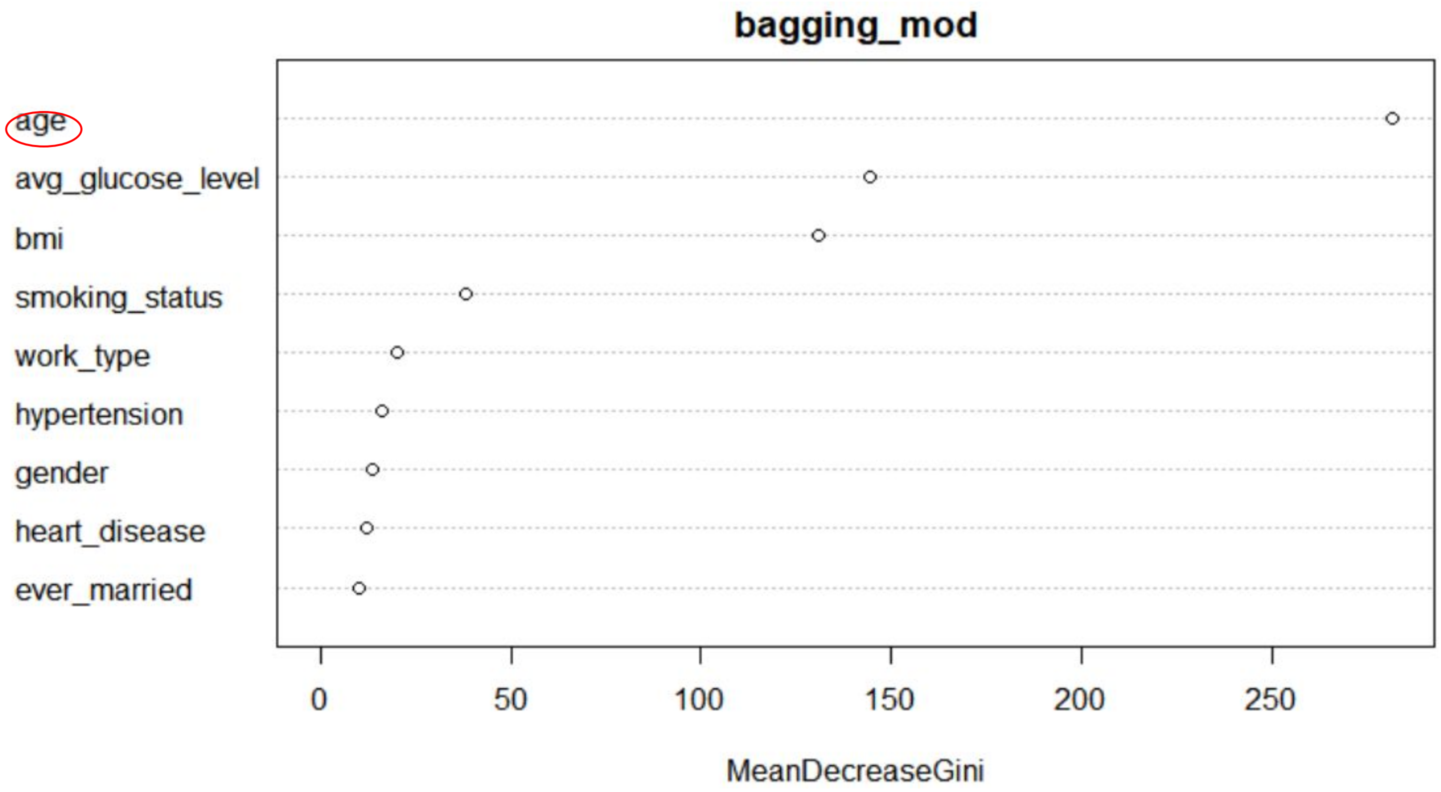


Bagging

Confusion Matrix

Actual	Non-Stroke	Stroke
Non-Stroke	14865	3175
Stroke	29	40

- Misclassification Rate=0.18
- Precision= 0.012
- Recall= 0.58



Model Comparison

KNN

Misclassification Rate: 0.112

Precision: 0.112

Recall: 0.883

Logistic Regression

Misclassification Rate: 0.165

Precision: 0.069

Recall: 0.064

Tree

Misclassification Rate: 0.20

Precision: 0.014

Recall: 0.67

Boosting

Misclassification Rate: 0.13

Precision: 0.065

Recall: 0.49

Random Forest

Misclassification Rate: 0.15

Precision: 0.012

Recall: 0.55

Bagging

Misclassification Rate: 0.18

Precision: 0.012

Recall: 0.58

Questions?