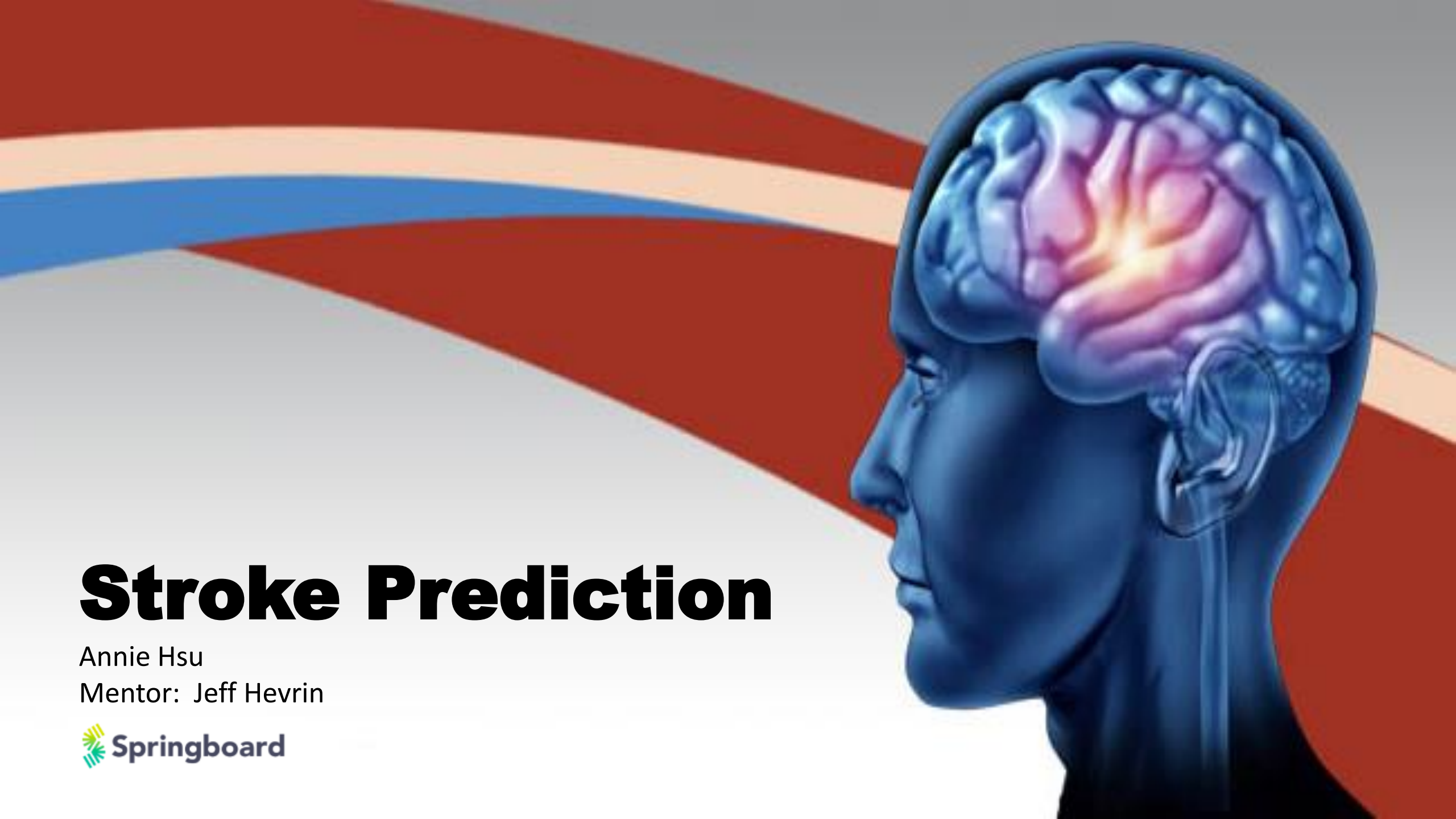


Stroke Prediction

Annie Hsu
Mentor: Jeff Hevrin

Springboard

Stroke

- No. 5 cause of death
- Leading cause of long-term disability
- Every year 800,000 people have stroke
- 25% of cases are recurrent

What factors increase the risk?

Can we predict the likelihood of having a stroke?

Dataset

- Dataset was taken from Kaggle.com
- 5110 entries of patient information
- 11 features

Features

- ID
- Gender
- Age
- Hypertension
- Heart disease
- Ever married
- Work type
- Residence type
- Avg. glucose level
- BMI
- Smoking status

Data Wrangling

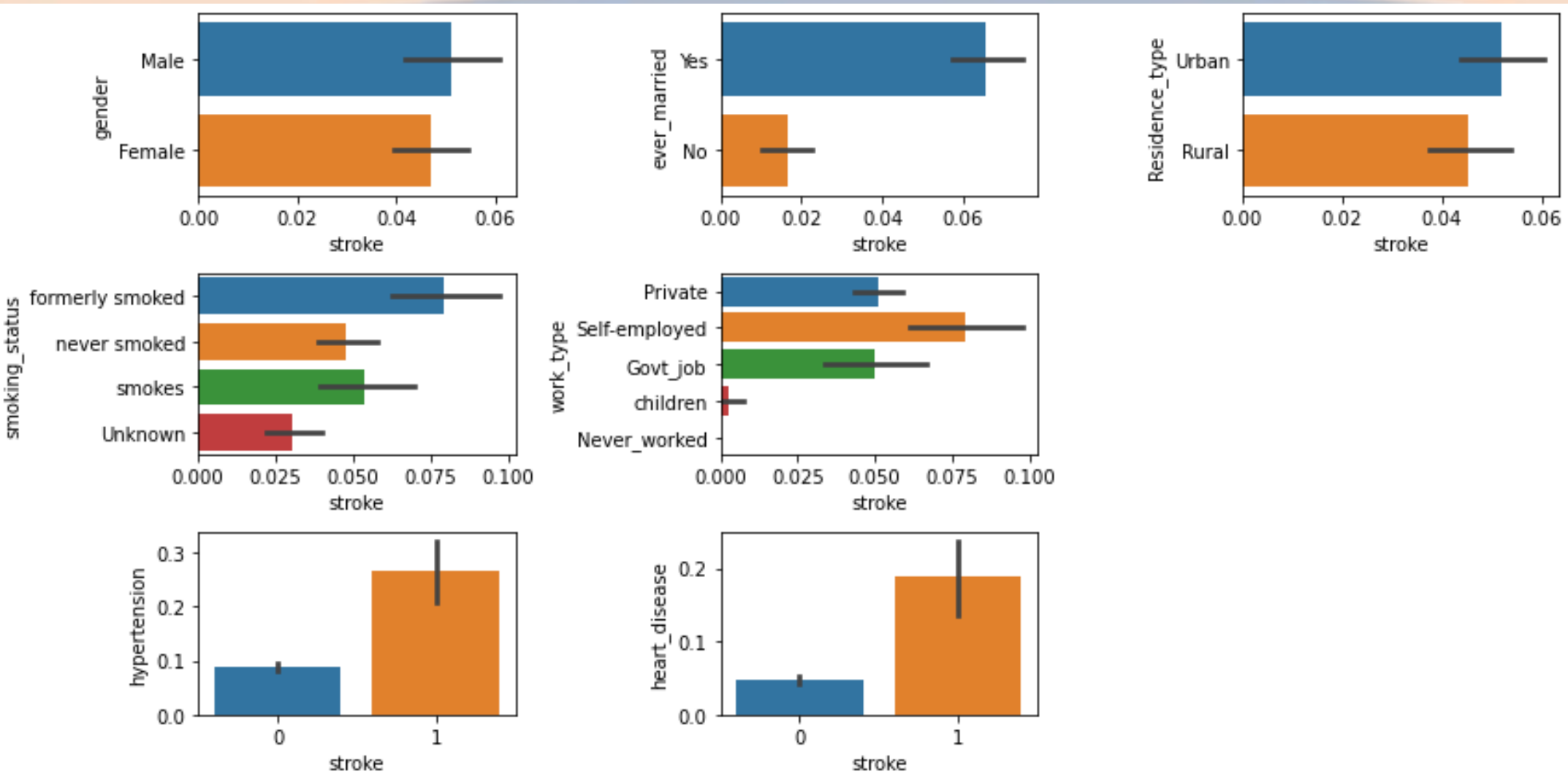
- **ID (dropped)**
- **Gender (dropped “Other”)**
- Age
- Hypertension
- Heart disease
- Ever married
- Work type
- Residence type
- Avg. glucose level
- **BMI (~4% missing, filled with median)**
- Smoking status

Exploratory Data Analysis

Categorical features

- Gender
- Hypertension
- Heart disease
- Ever married
- Work type
- Residence type
- Smoking status

EDA (Categorical features)

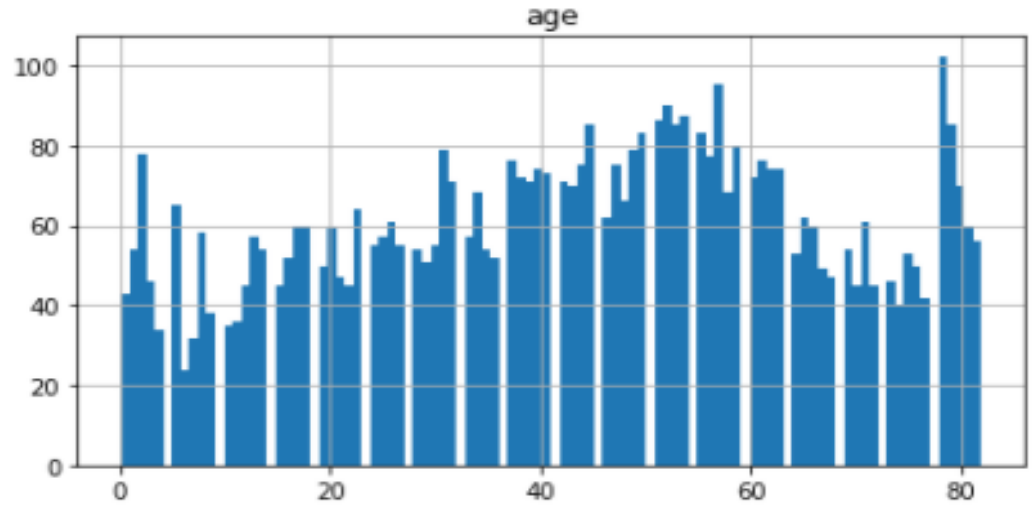


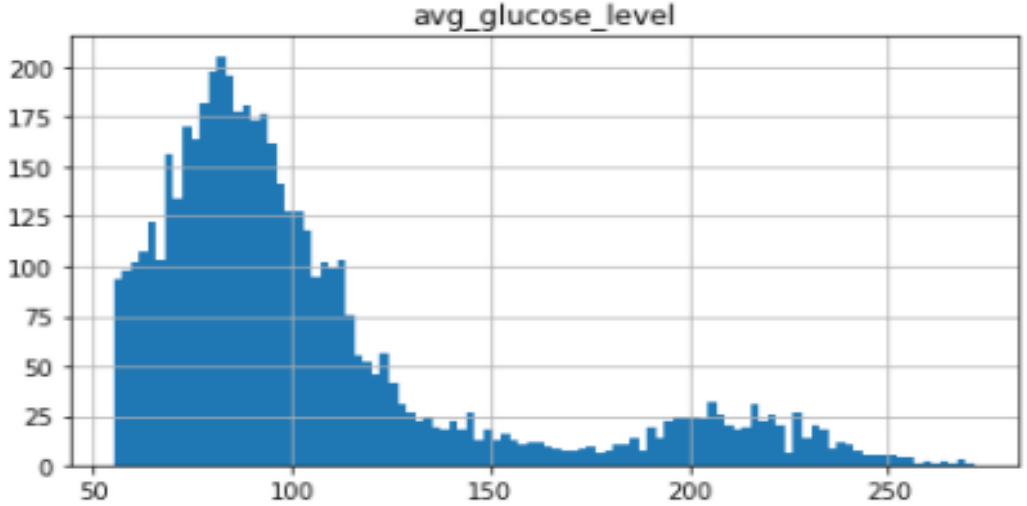
Exploratory Data Analysis

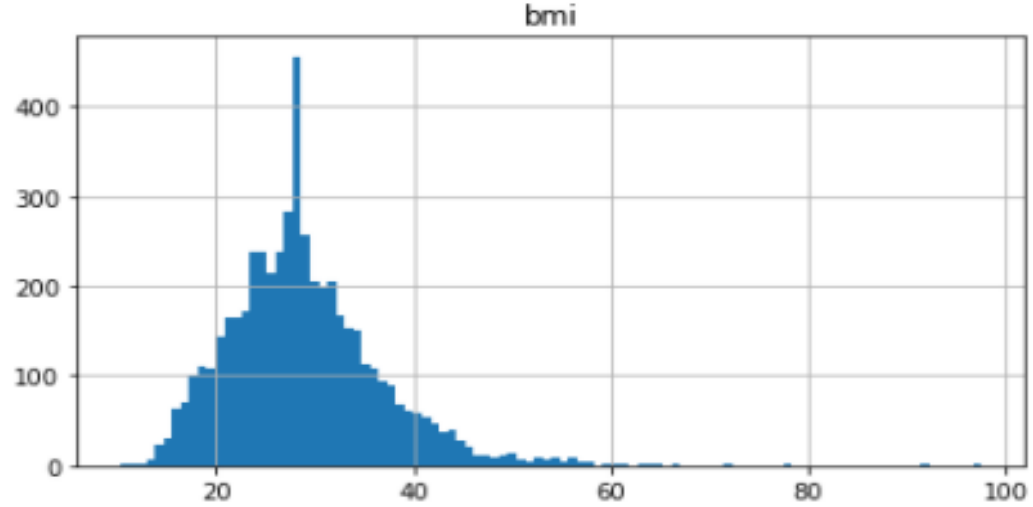
Numerical features

- Age
- BMI
- Avg. glucose level

EDA (Numerical features)

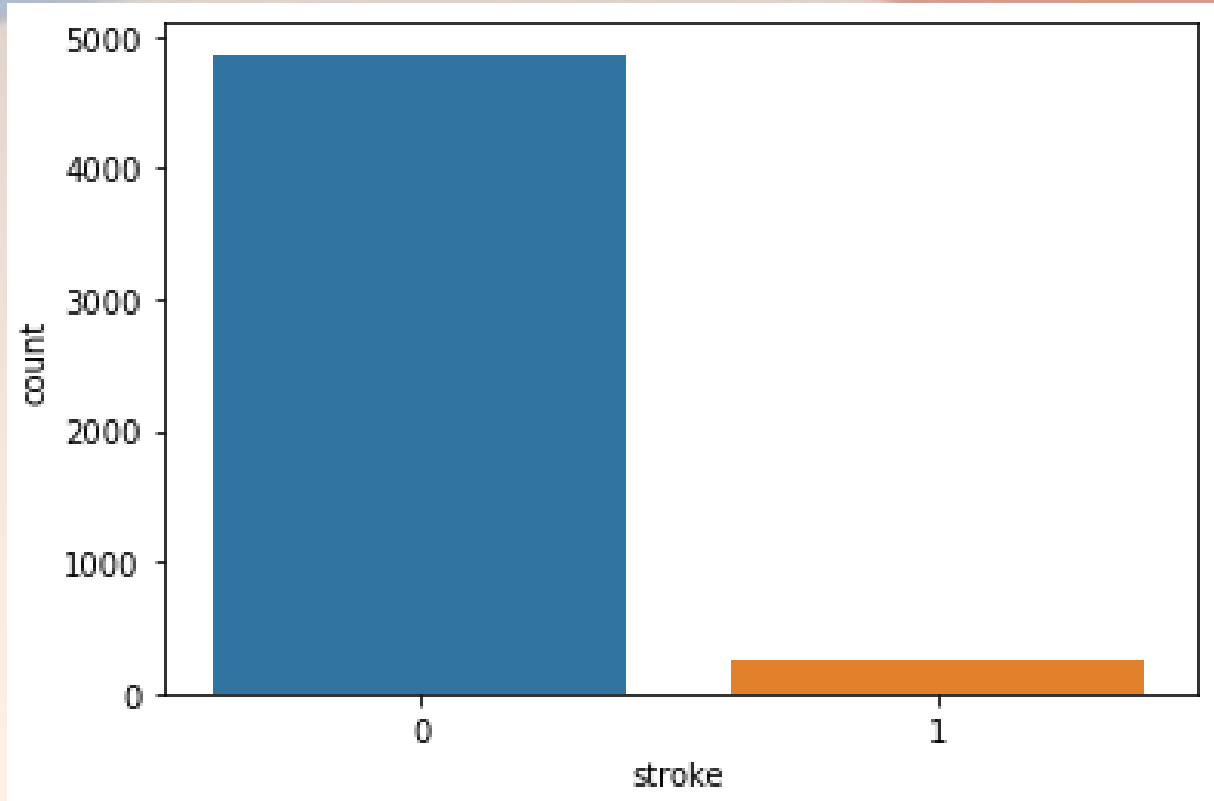






Exploratory Data Analysis

Target - ***Stroke***



Data Modeling

- Train-test split: 80-20

Predictive models

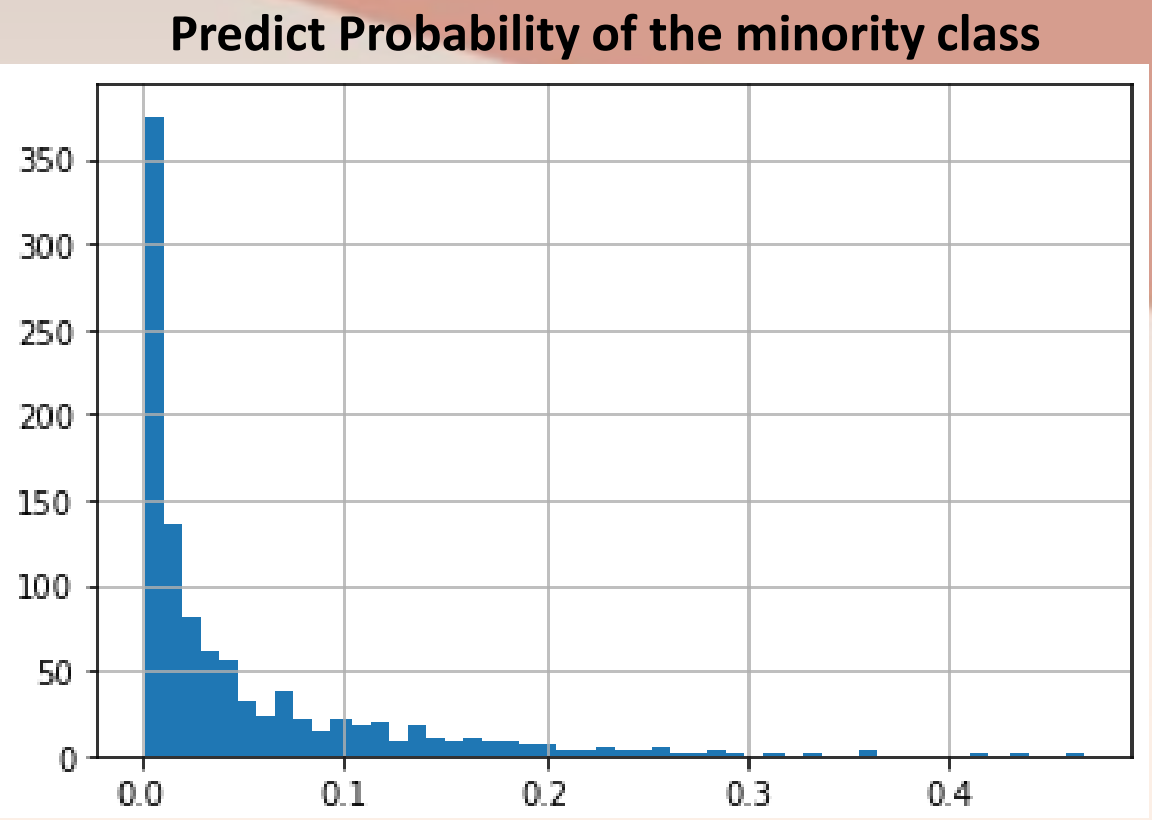
- Logistic Regression
- K-Nearest Neighbors
- Decision Tree
- Random Forest
- Gradient Boost
- XGBoost

Findings

First model: Logistic Regression
Train accuracy: 95.2%
Test accuracy: 95.1%
Precision rate: **0%**
Recall rate: **0%**

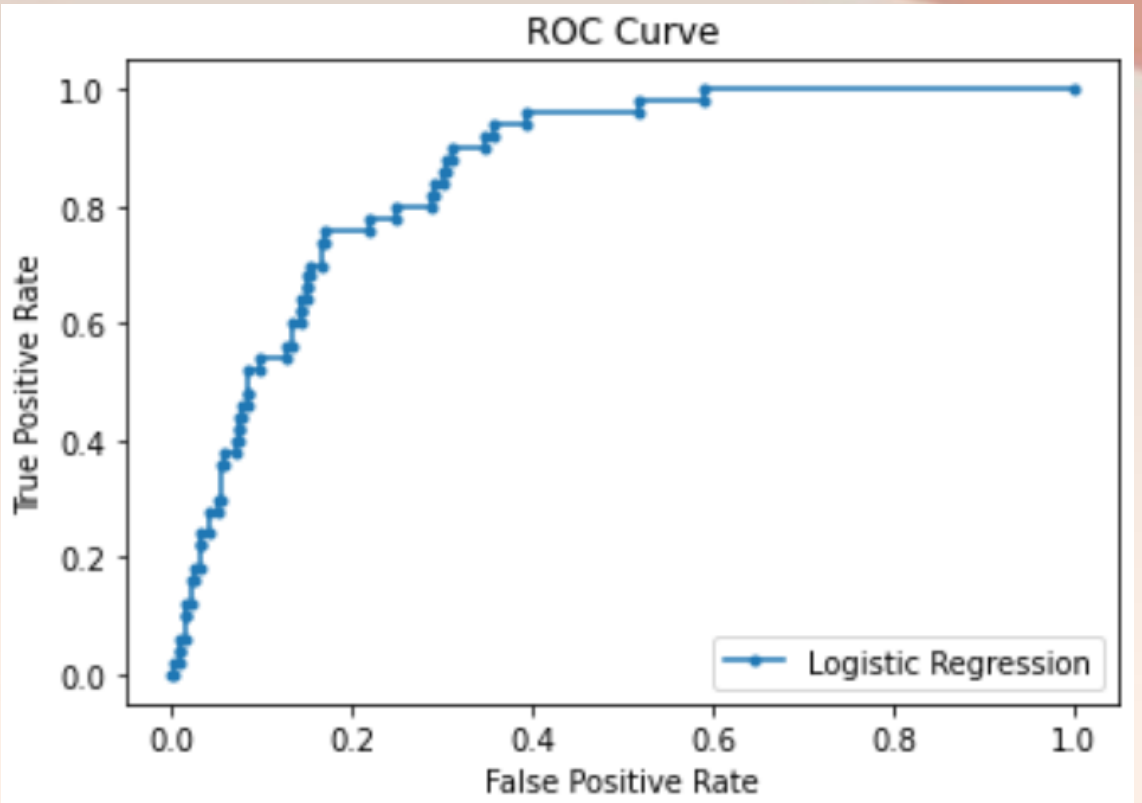
Model Diagnostics

First model: Logistic Regression
Train accuracy: 95.2%
Test accuracy: 95.1%
Precision rate: **0%**
Recall rate: **0%**



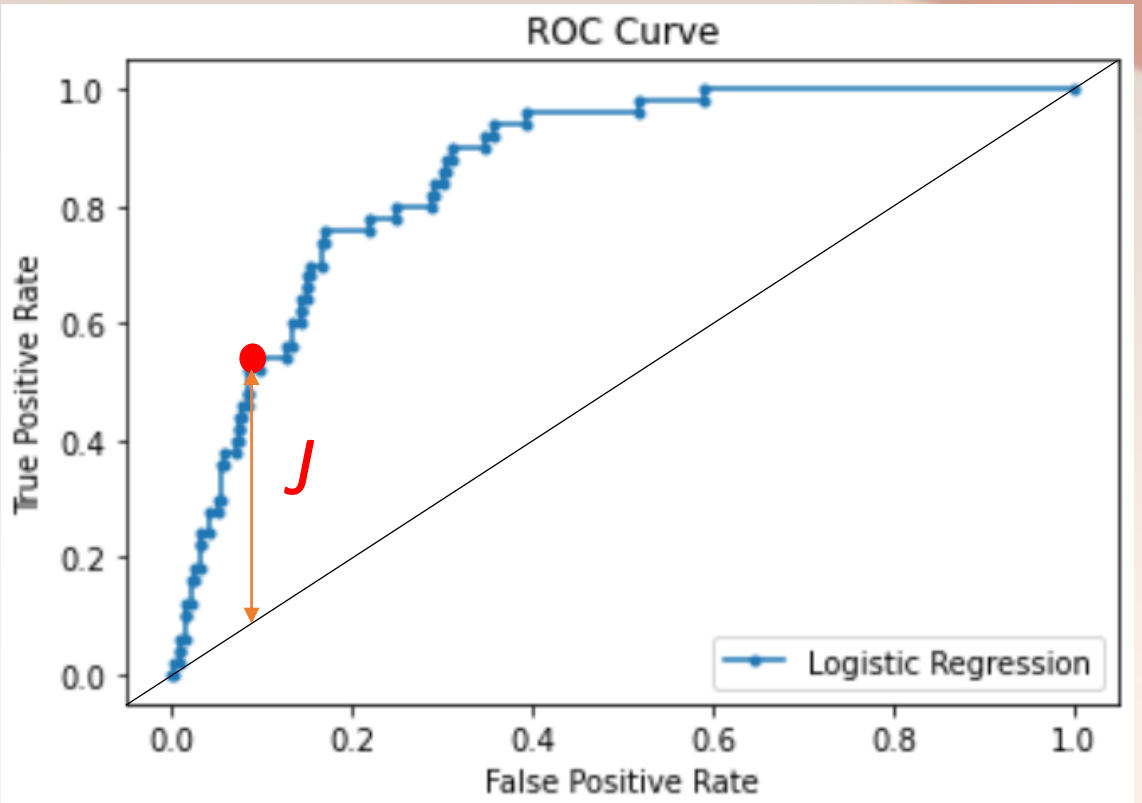
Moving-Threshold method

ROC curve - Performance of the classifier at all thresholds



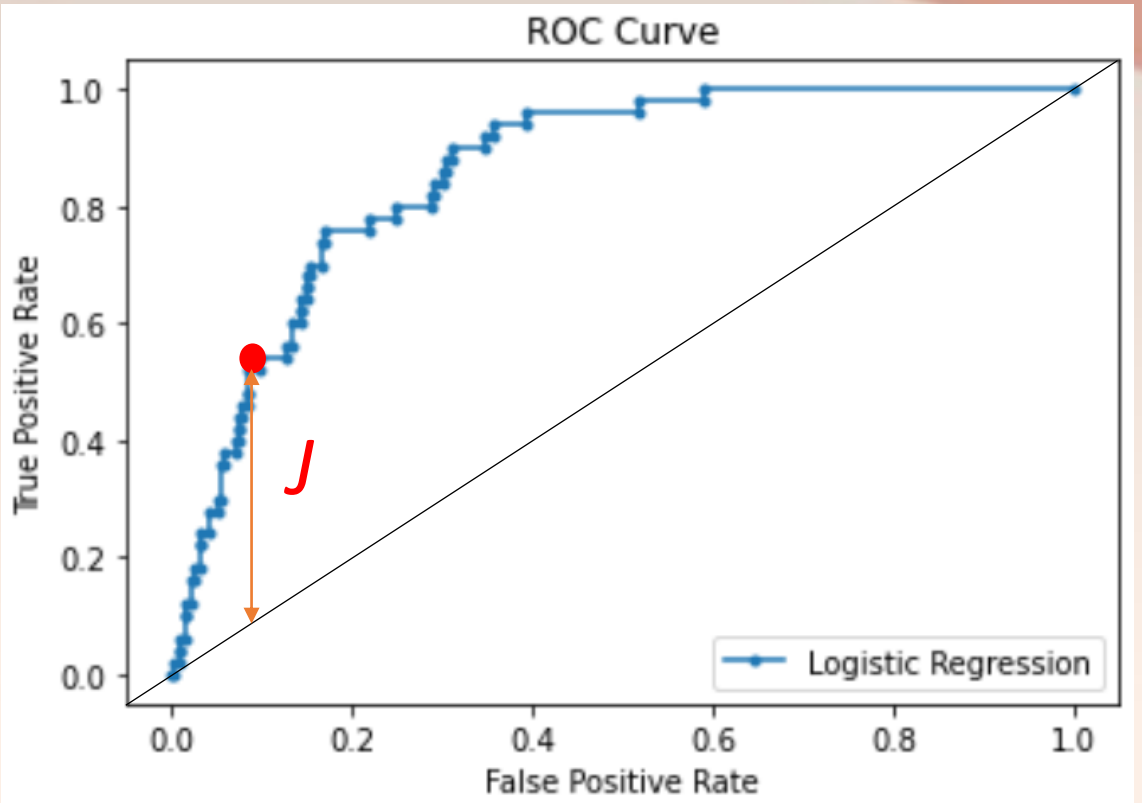
Moving-Threshold method

Find the optimal threshold using Youden's J statistic



Moving-Threshold method

Find the optimal threshold using Youden's J statistic



Optimal threshold = 0.08
Recall rate = 74%
AUC ROC = 0.86

Model Evaluation Metrics

- Recall rate
- AUC ROC
- Diagnostic Odds Ratio (DOR)

$$\text{DOR} = \frac{\text{Probability of detection}}{\text{Probability of false alarm}} = \frac{\text{True positive / False Positive}}{\text{False Negative / True Negative}}$$

Models Performance

Model	ROC AUC score	DOR score
Tuned Gradient Boosting	0.876	19.483
Tuned Logistic Regression	0.859	16.506
Tuned Random Forest	0.858	15.833
Gradient Boosting	0.862	15.081
Logistic Regression	0.859	14.023
Random Forest	0.790	8.023
XGBoost	0.809	4.985
K-Neighbors	0.709	4.358
Decision Tree	0.568	NaN