Stroke Prediction

Data Science Capstone Project, December 17, 2022

The Problem Statement

According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally.

Is it possible to predict whether or not an individual will suffer a stroke?

What factors affect stroke occurrence and can we predict likelihood of a stroke?

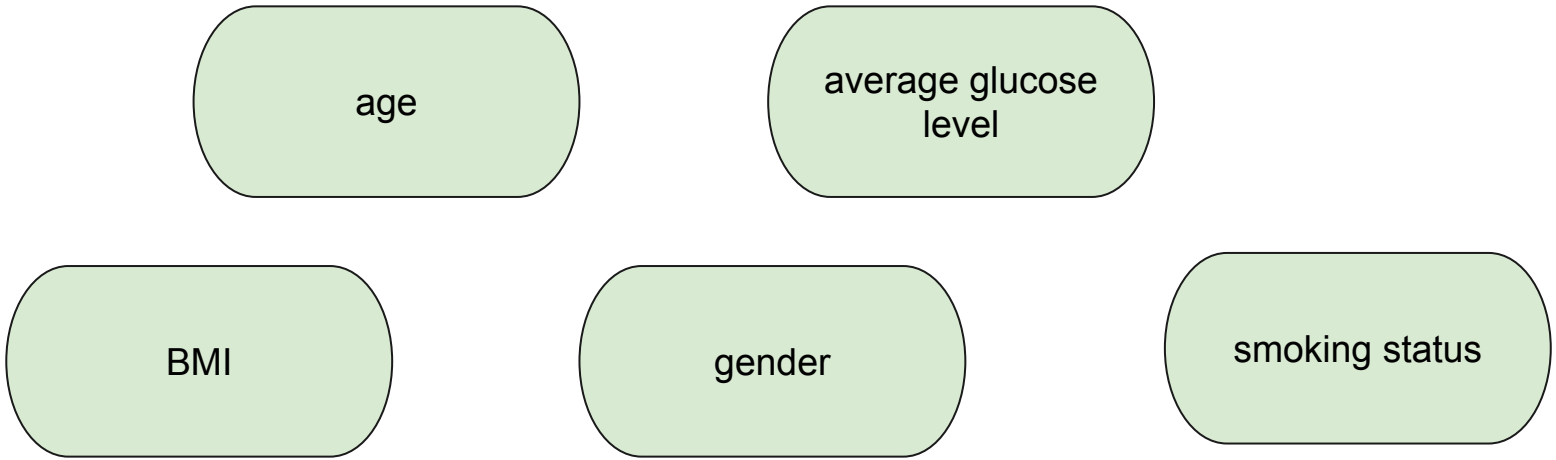
Who might care?

Hospitals

Patients

Healthcare Providers

What factors might affect someone having a stroke?



Feature list and target variable (Kaggle dataset)

1. id: unique identifier

2. gender: "Male", "Female" or "Other"

3. age: age of the patient

4. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5. heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6. ever_married: "No" or "Yes"

7. work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8. Residence_type: "Rural" or "Urban"

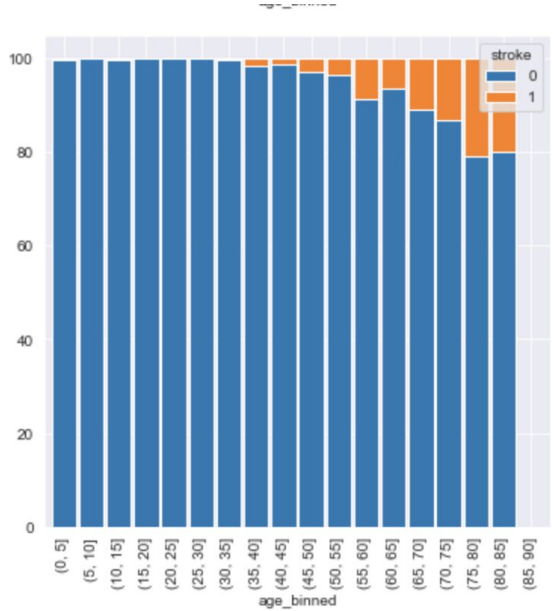
9. avg_glucose_level: average glucose level in blood

10. bmi: body mass index

11. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12. stroke: 1 if the patient had a stroke or 0 if not

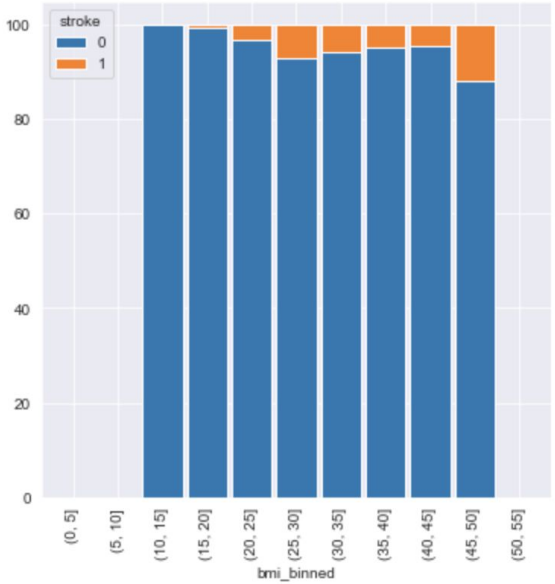
Data Exploration - Age and Stroke



The bar chart on the left shows the stroke occurrence among different ages.

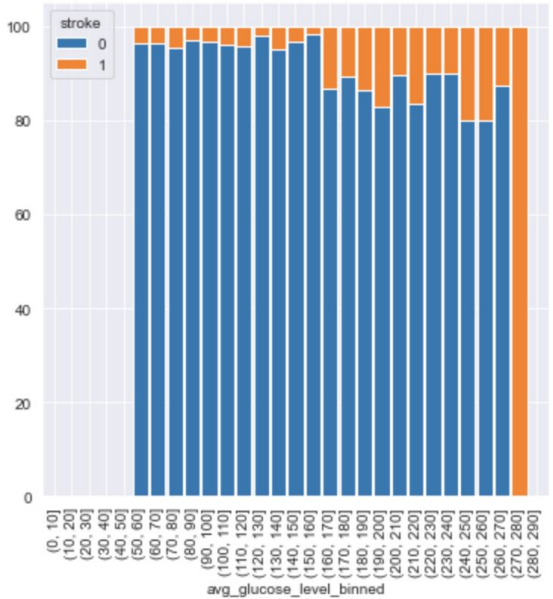
Older patients are more likely to suffer from a stroke.

Data Exploration - BMI and Stroke



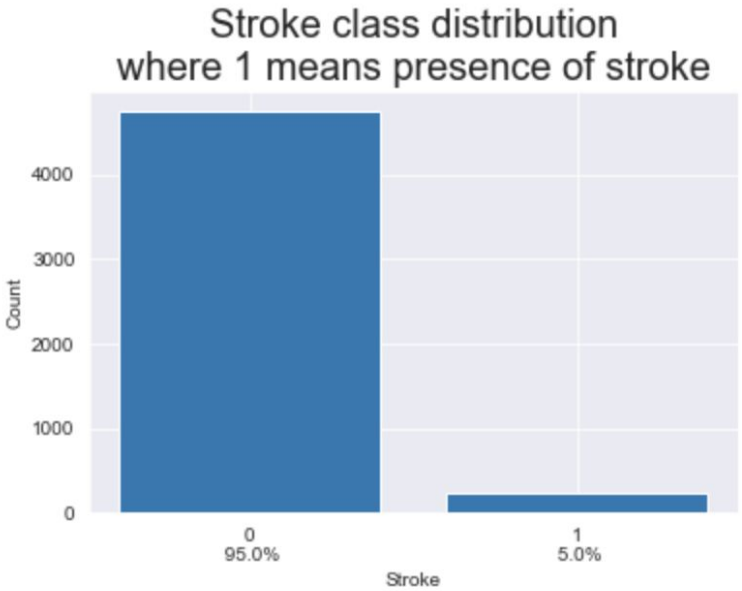
Higher BMI was associated with a higher chance of stroke.

Data Exploration - Average Glucose Level and Stroke



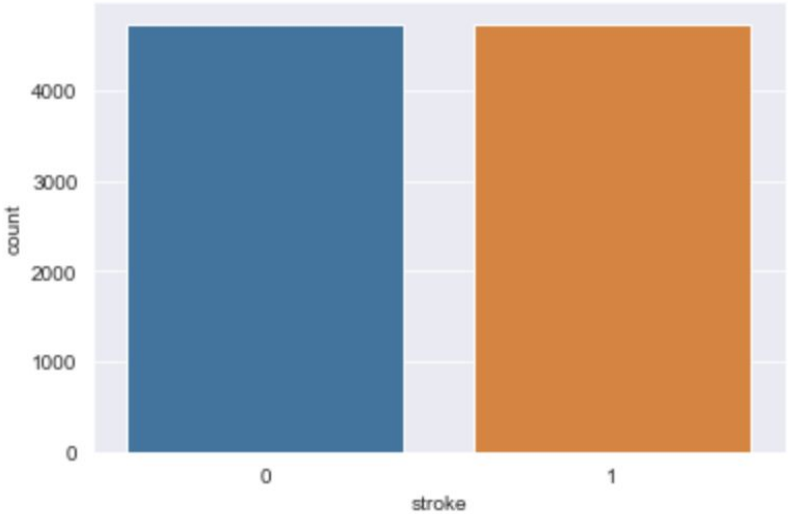
Higher average glucose level was also associated with a higher likelihood of stroke.

Imbalanced Dataset



Dataset is imbalanced with stroke patients accounting for a minority of cases.

SMOTE



We used SMOTE to create balanced stroke / no stroke groups.

Data Pre-processing

Data was divided into 20% test and 80% training data.

The numerical columns were scaled using StandardScaler.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
6331	1	0.027994	0	0	0	1	1	-0.575887	0.409208	0
6416	0	-0.369319	0	0	1	2	0	-0.737635	0.289460	1
8419	0	0.821503	1	0	1	3	1	1.430906	0.598884	2
7254	0	1.134569	0	0	1	2	1	-0.806727	-0.781485	0
1495	0	-0.045599	0	0	1	2	1	1.658289	1.644859	2

Modeling

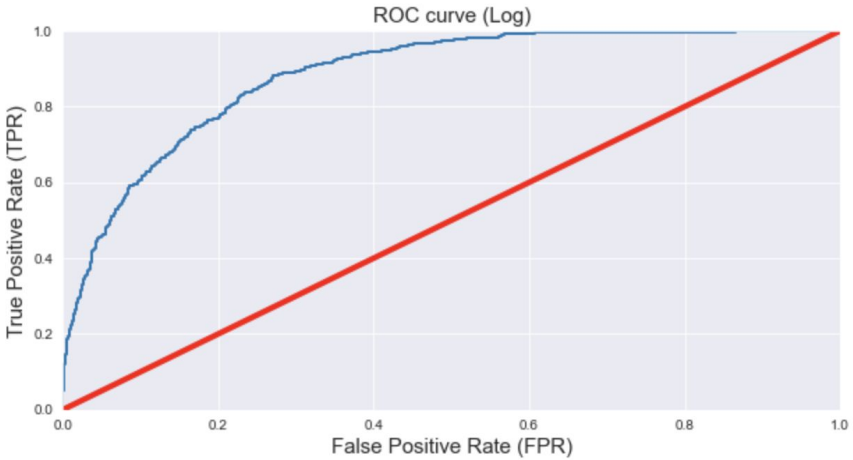
Three models were used on this dataset. All models used RandomizedSearchCV to find optimal parameters.

- Logistic Regression

- Support Vector Machine

- Random Forest

Logistic Regression

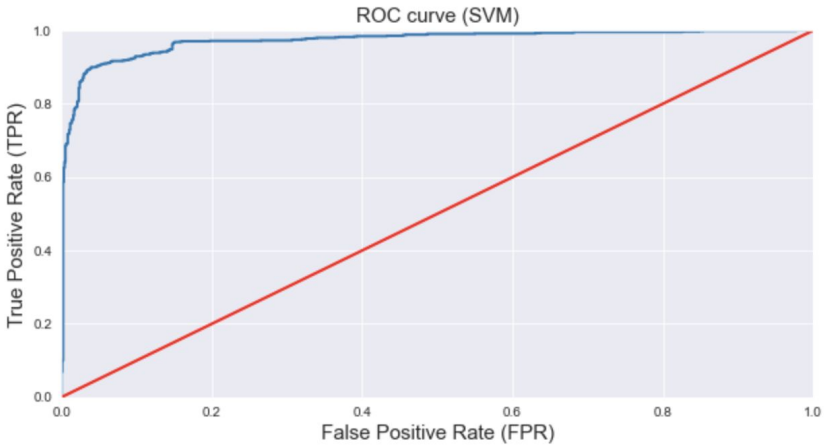


Logistic Regression model showed the following:

- 79% accuracy (TP + TN) / (TP + FP + TN + FN)

- 81% recall for stroke patients (TP / (TP + FN)

Support Vector Machine (SVM)

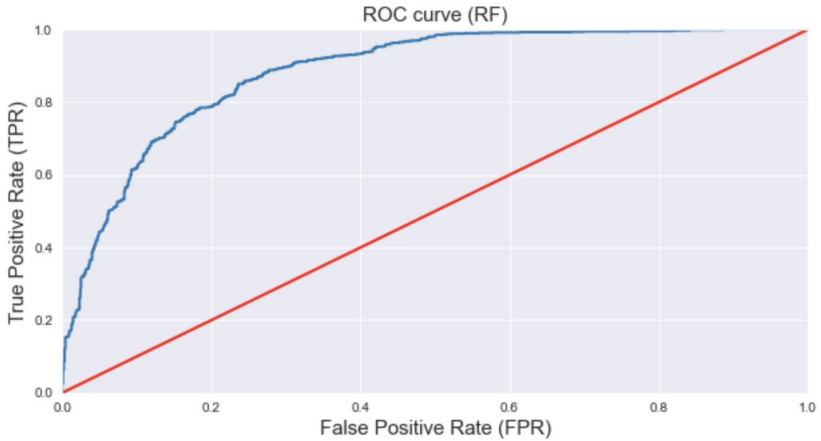


SVM model showed the following:

- 93% accuracy (TP + TN) / (TP + FP + TN + FN)

- 90% recall for stroke patients (TP / (TP + FN)

Random Forest



Random Forest model showed:

- 80% accuracy (TP + TN) / (TP + FP + TN + FN)

- 91% recall for stroke patients (TP / (TP + FN)

Model Comparison

The **SVM model** gave the best accuracy and recall. We care about recall (TP / TP + FN) because we don't want to misdiagnose people who are likely to suffer from stroke as negative.

Algorithm	ROC-AUC train score	ROC-AUC test score
Logistic Regression	0.888266	0.881759
SVM	0.957949	0.906161
Random Forest	0.894338	0.878049