Using Machine Learning to Predict Stroke*

*A stroke, sometimes called a brain attack, occurs when something blocks blood supply to part of the brain, or when a blood vessel in the brain bursts--causing parts of the brain to become damaged or die.
https://www.cdc.gov/stroke/about.htm

Aaron Pisacane

Sarah Caldwell

Ariel Jones

Seamus Murphy

Sulman Choudhary

Stroke Prediction Data Set

Did you know...

Stroke is a leading cause of serious long-term disability

Someone in the U.S. has a stroke every 40 seconds

In 2020, 1 in 6 heart-disease related deaths were due to stroke

Predicting a Stroke

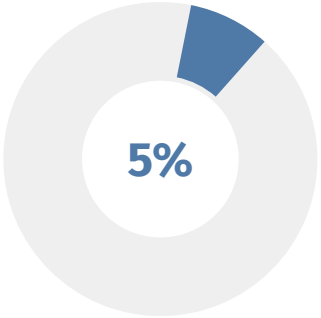
We retrieved a dataset from Kaggle that listed what we believe to be the key features to predict a stroke.

These features are:

- Gender
- Age
- Heart disease status
- Marital status
- Work type
- Residence type
- Average Glucose
- BMI
- Smoking status
- Hypertension

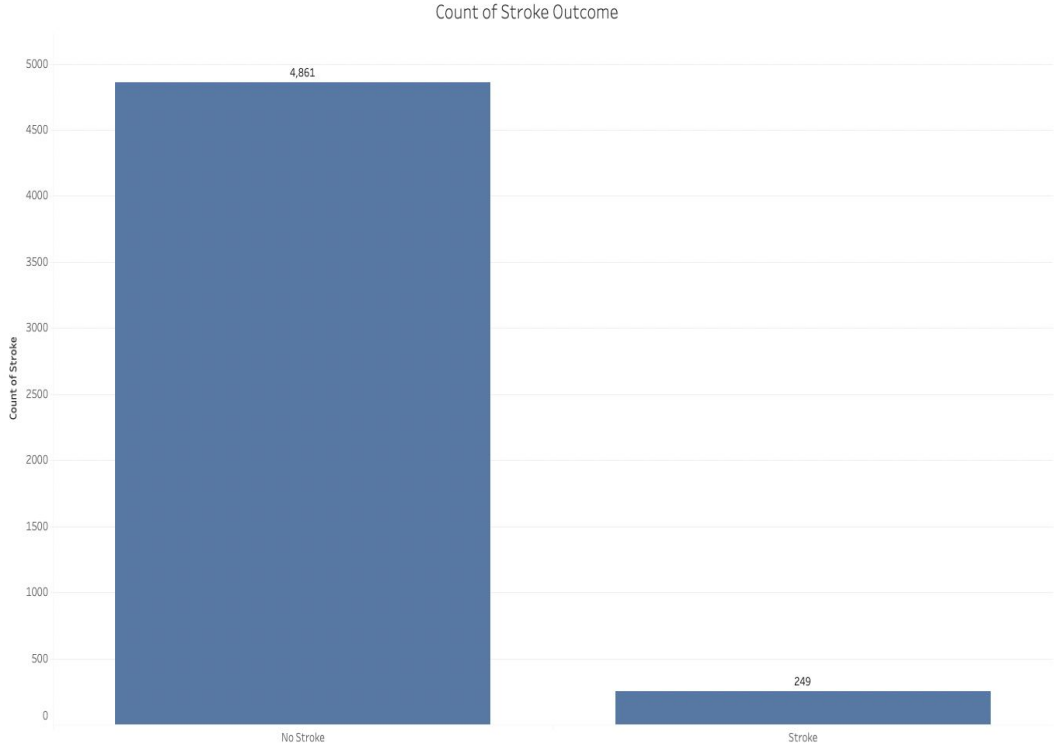
The top three most important features were ***AGE, GLUCOSE LEVEL, AND BMI***.

The Data

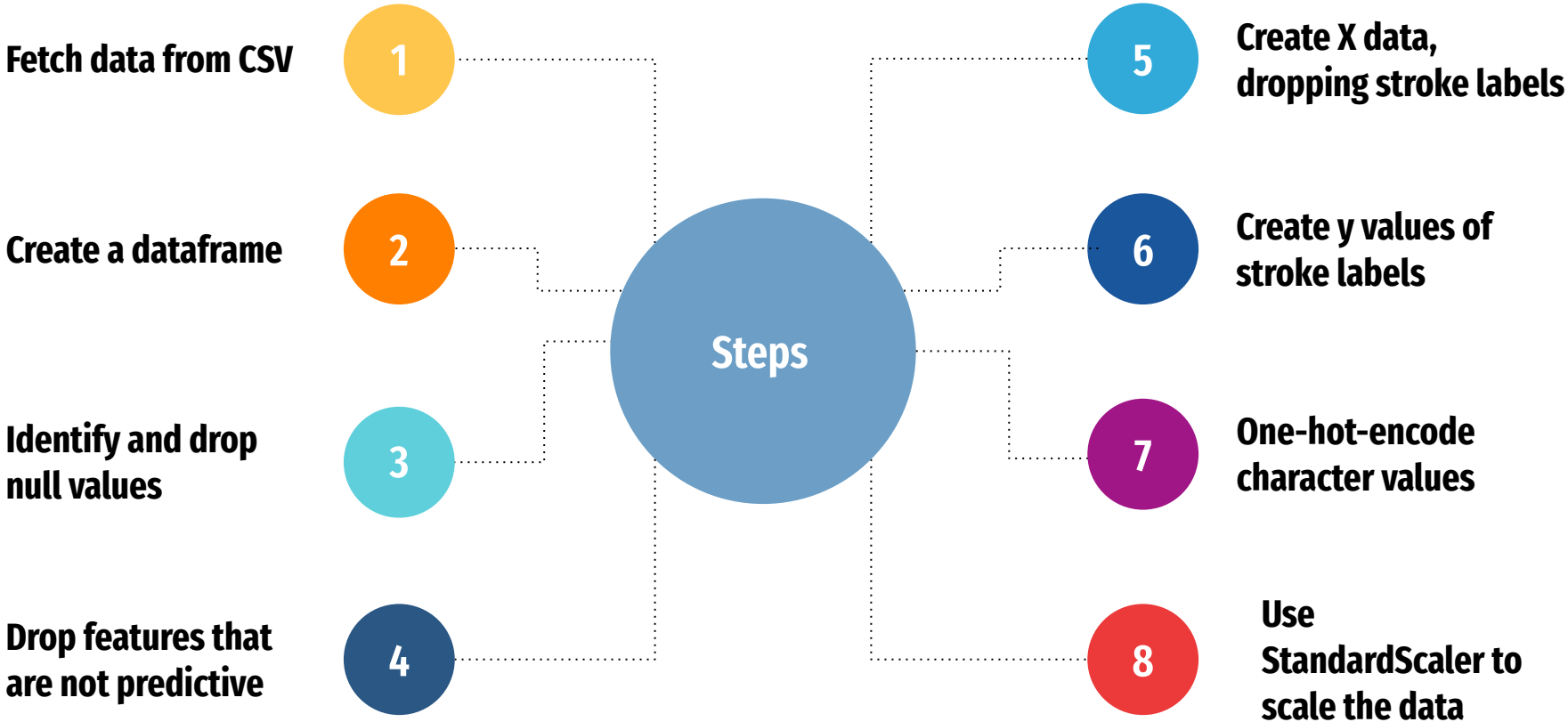


Distribution of Outcomes

Only approximately 5% of participants in dataset had a stroke



Preprocessing



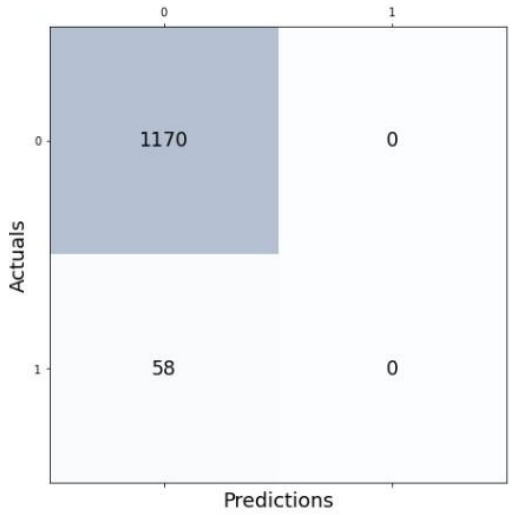
Logistic Regression Model

Results

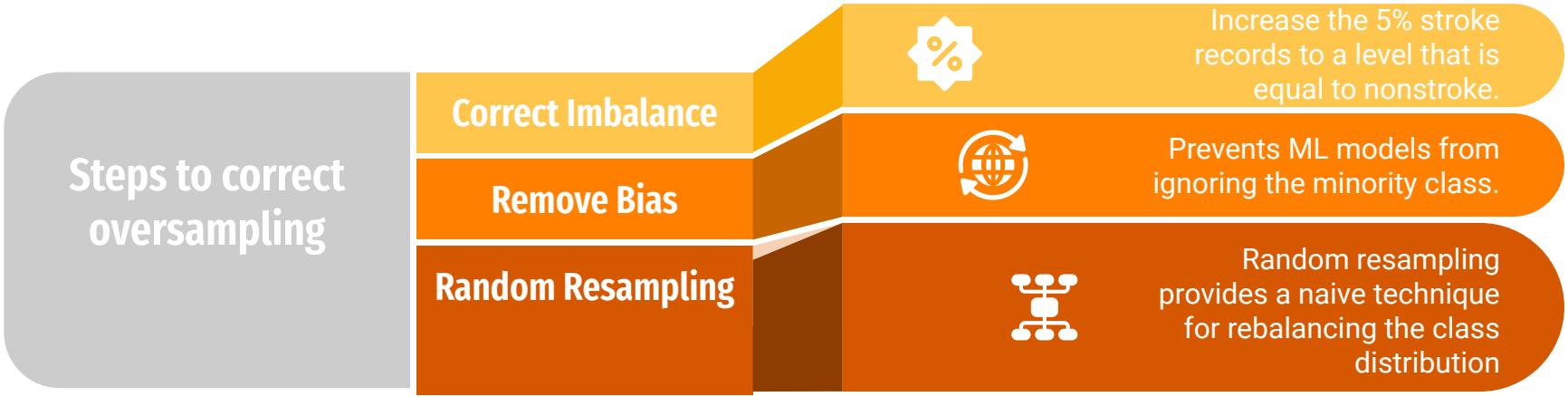
Training Score=0.959
Testing Score Score=0.953
Accuracy=0.95

AUC =0.5

Confusion Matrix



Oversampling




```
# Increase the 5% (250 records) stroke (1) sample to equal the non-stroke(0) through RandomOverSampler

from imblearn.over_sampling import RandomOverSampler

oversample = RandomOverSampler(sampling_strategy='minority')
X_over, y_over = oversample.fit_resample(X_dummies, y)
```

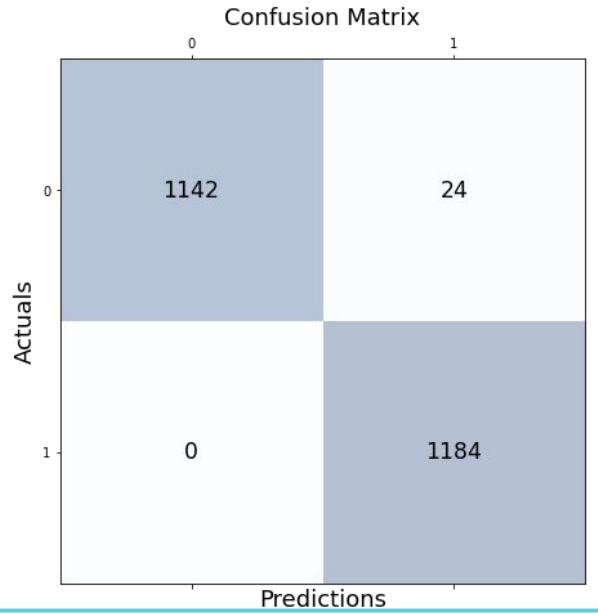
Random Forest Model

Results

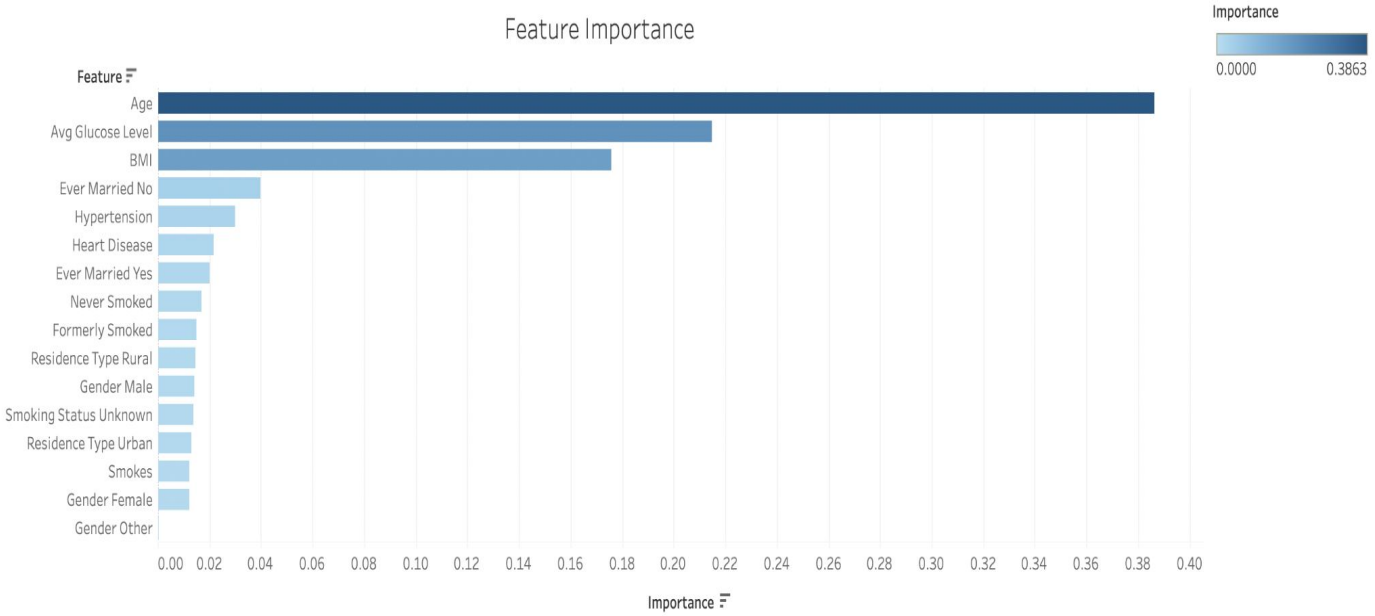
Training Score=0.988
Testing Score Score=0.988
Accuracy=0.99

AUC =0.988

Confusion Matrix



Feature Importance



Age

Importance: 0.3863

Avg. Glucose

Importance: 0.2149

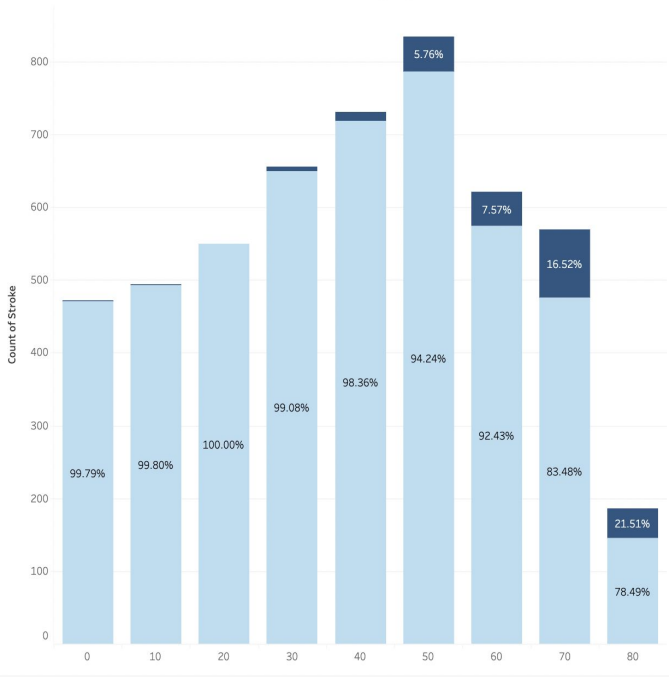
Body Mass Index (BMI)

Importance: 0.1759

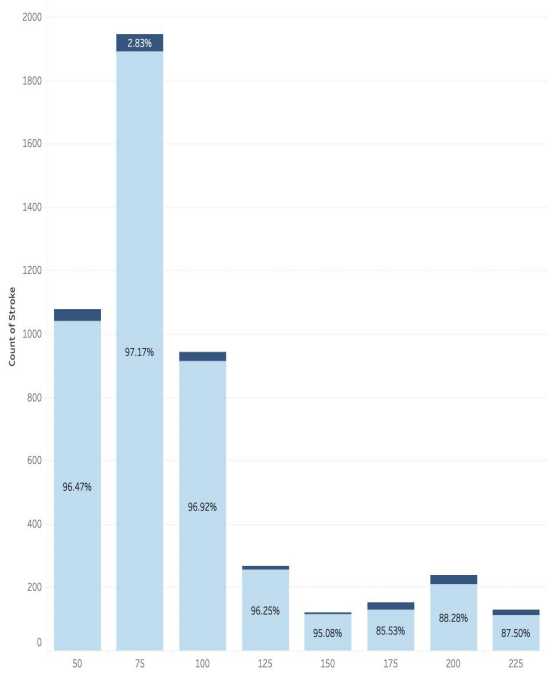
Important Features

Outcome
No Stroke
Stroke

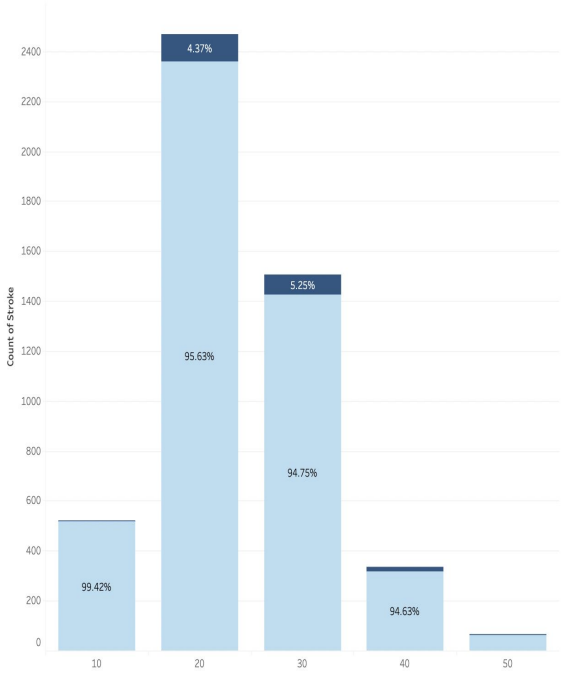
Age



Avg. Glucose Level



Body Mass Index (BMI)



Conclusion

We found that a Random Forest Model was most effective at predicting stroke. Factors such as age, BMI, and blood glucose level relate heavily to whether someone is at increased risk of stroke.

We like to think of medicine as an exact science. It is clear that Machine Learning can help us pull meaning from large datasets, identify important factors, and spot similarities. In turn, we can find connections between a current patient and the vast amount of data from previous patients/study subjects and provide better, life-saving care.

Next Steps

To improve accuracy and usability of our model, we could:

Improve the balance of the dataset:

Collect more data from stroke patients or train a model from a dataset that already has more balanced data

Test models that can further classify percentage chance of a stroke

Propose further study into the links between stroke and the factors of BMI, blood sugar, and age

Questions