Sherbourne Stroke Prediction

Ray Sherbourne

4/6/2021

Project Overview

Strokes account for 11% of global deaths according to the WHO (World Health Organization). The goal of this project is to develop a predictive model that can be used as an aid to identify at-risk individuals. The data is downloaded from kaggle.com. The author, fedesoriano, states the source is confidential.

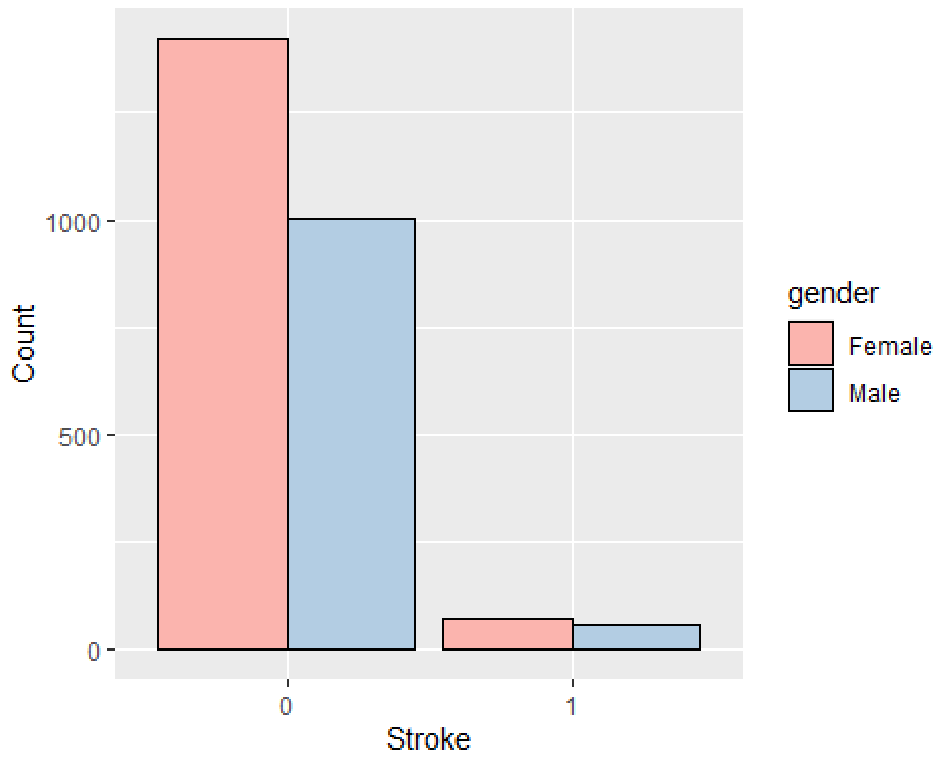
The data set contains 5110 observations and 12 predictors including the binary outcome indicating stroke. The predictors also include an anonymous id, and information for gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index, and smoking status. No other background information is given about the data.

After a validation set has been partitioned from the data, the remaining data will be cleaned, explored, and used to build several models. The best model will be tested against the validation set, and the resulting scores will determine if the goal of this project succeeded. The main score that will be used to grade performance will be the **F1 score**. The F1 score is a harmonic average between precision and recall, and is a popular single number summary.

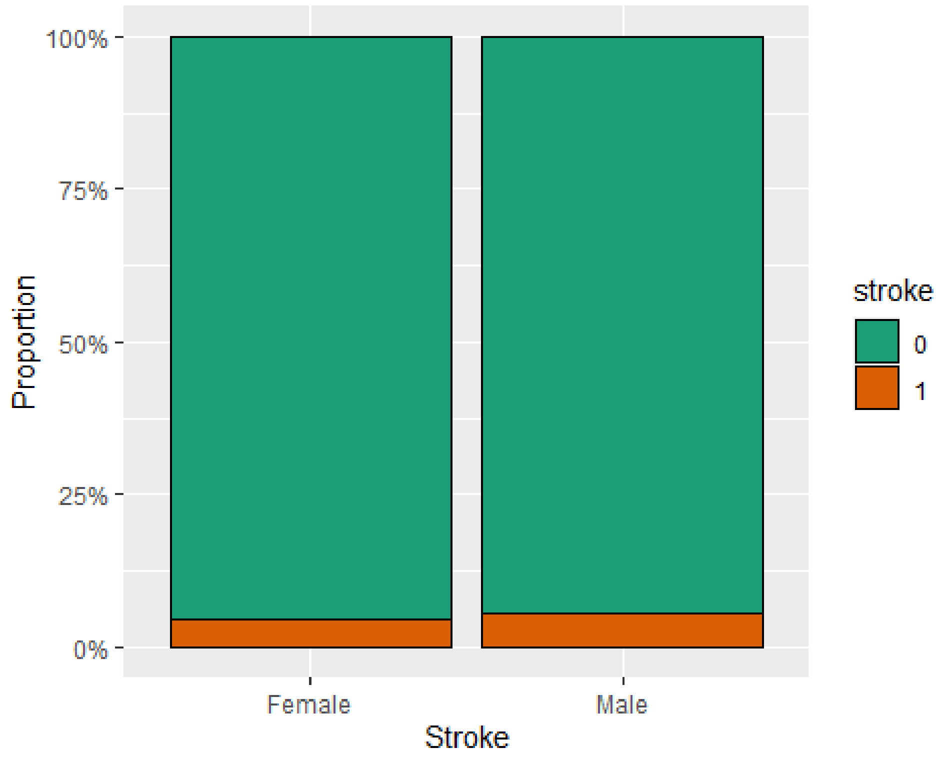
Method

The data was downloaded as a zip file from the kaggle website, extracted into the working directory of this project, and read into R. Because the data set is relatively small, only 10% was set aside as the validation set. The remaining data is cleaned by setting the predictor values from character types to either numeric values or factors. The bmi predictor has numerous NA values that will not work with the machine learning algorithms. For now these NA values are replaced with the average value for the sake of exploration.

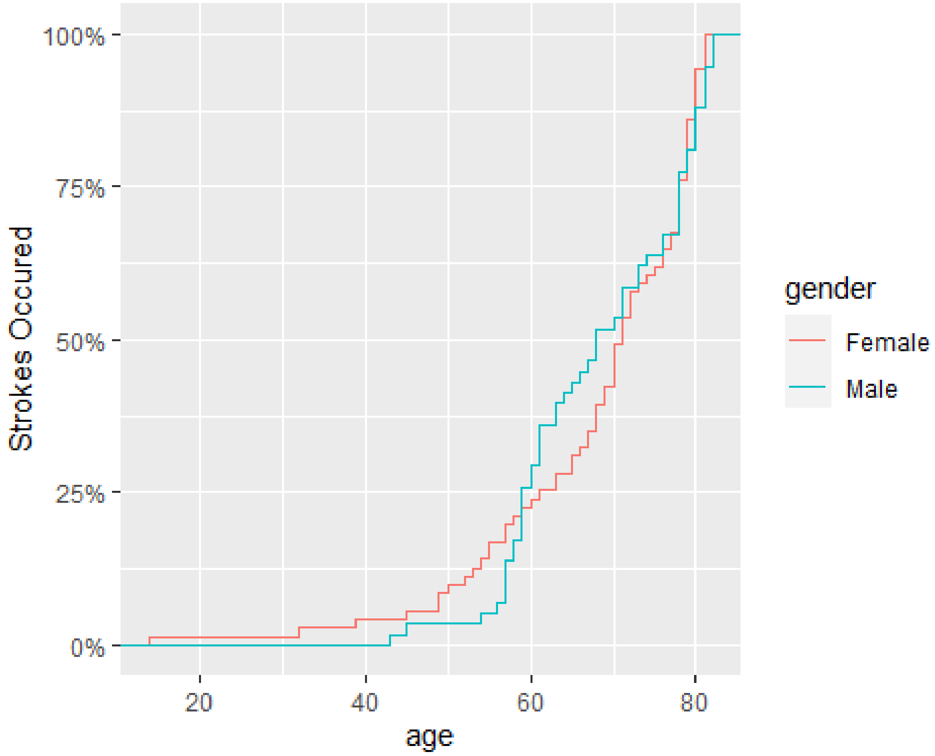
The first plot shows that the data contains more observations of females the proportional difference in strokes is not obvious however.



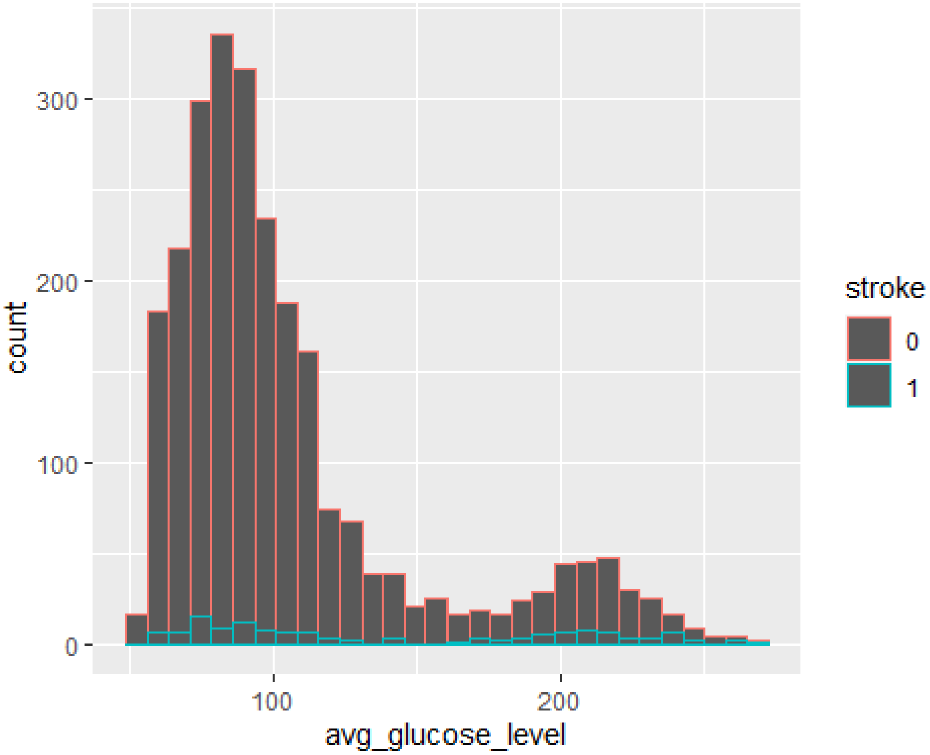
This plot shows that the rate of strokes is about the same between genders.



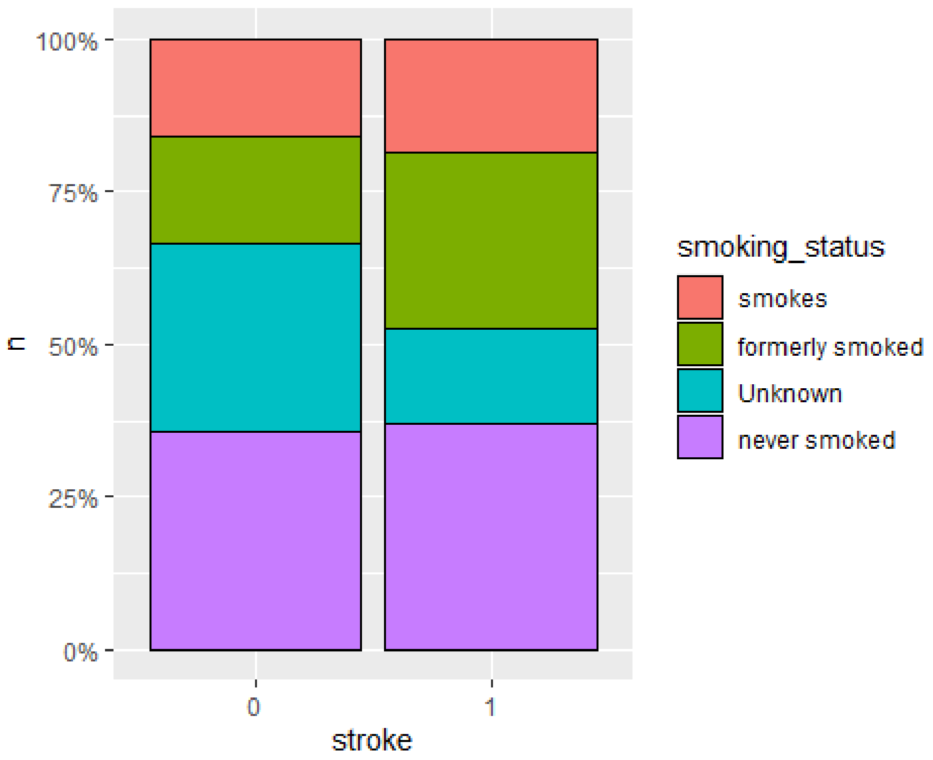
The next plot shows a cumulative distribution function of stroke patients for age and gender. Females have an exponential curve where the liklihood of a stroke increases rather smoothly. Men start off low and see a sharp increase in their mid to late 50's, overtaking the women until they even out in the early 70's.



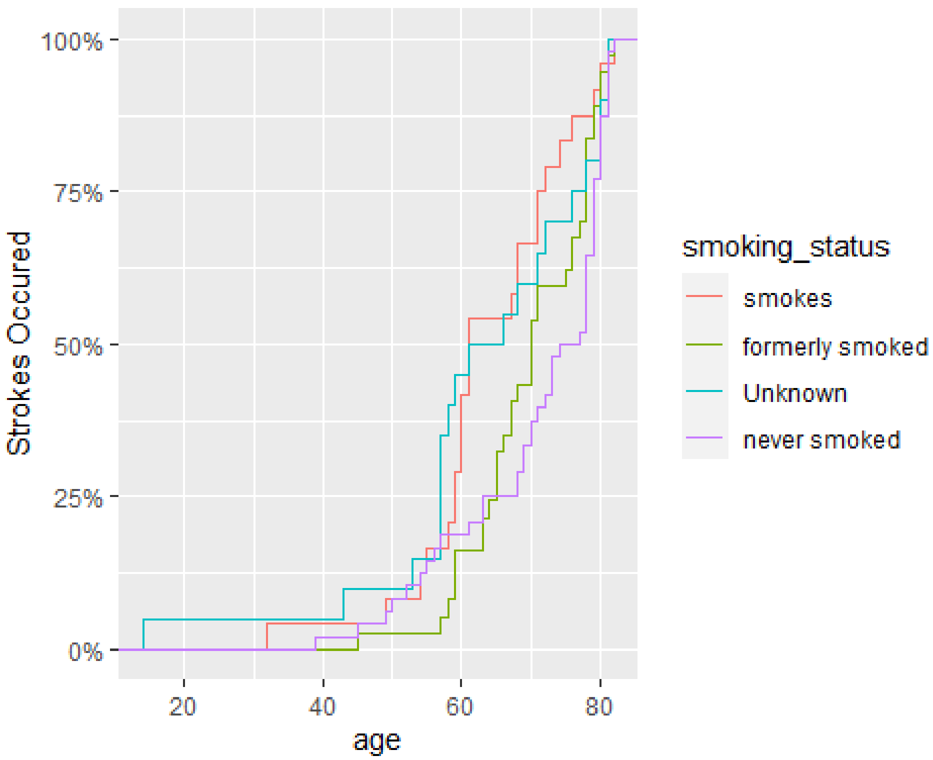
The next plot shows a histogram of average glucose level. The distribution is bimodal with the higher glucose levels being less prevalent and much higher percentage of strokes.



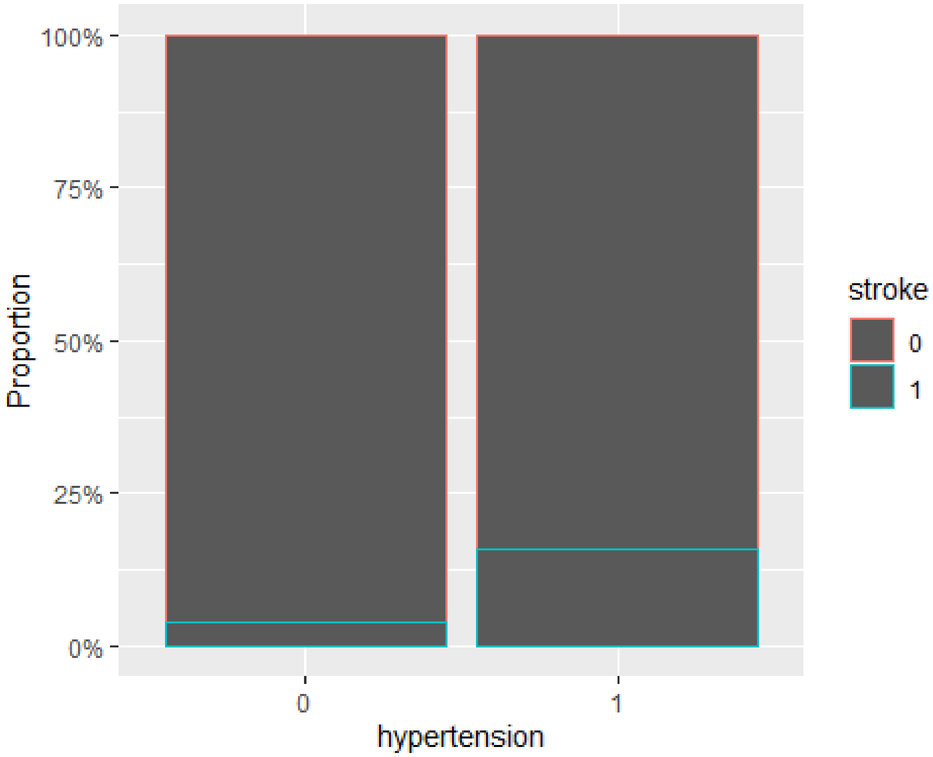
The next bar plot shows a bar of stroke positive and a bar of stroke negative cases, filled in with smoking status. Non smokers appear equally likely to get a stroke. The same is true for patients who smoke their entire life. This graph suggests that quitting increases the chances of getting a stroke.



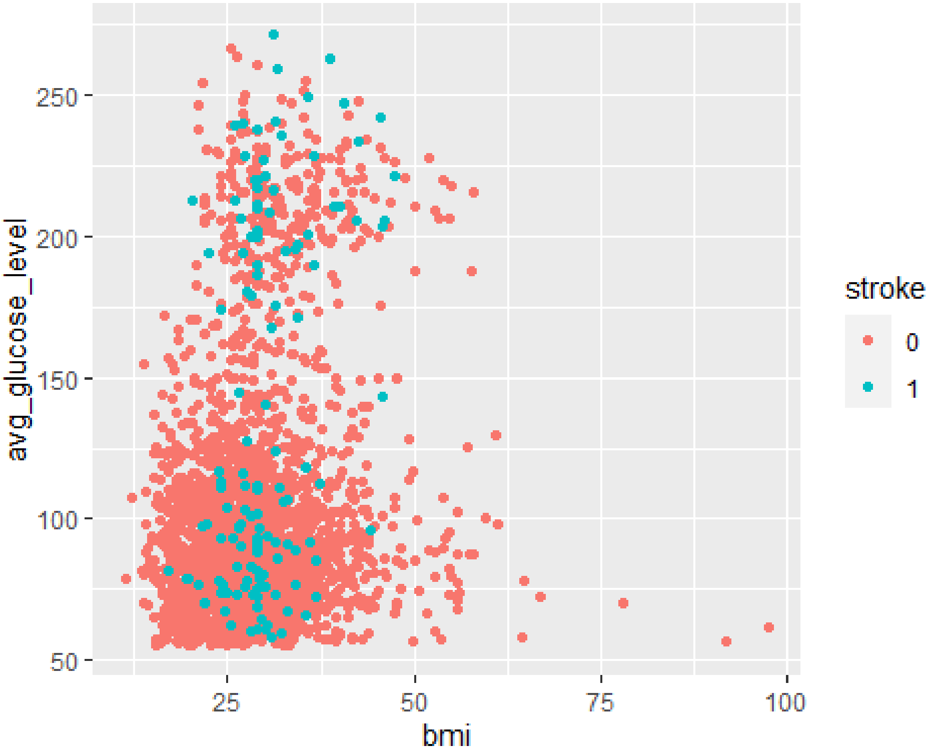
It is not clear if the stroke cases quit smoking for a few days and promptly relapsed, or if the non stroke patients quit smoking after only several years of smoking. There is no way to account for length of smoking abstinence with the given data, however the next plot shows a CDF of strokes and age for each of the smoking categories. This plot suggests that smokers get strokes earlier than non smokers, and former smokers are clearly having strokes at older ages than people who did not quit.



The next plot shows that hypertension is higher in stroke patients than non stroke patients.



The next plot trys to find a relationship between bmi and average glucose level. A relationship with bmi is wanted because there are so many NA values in the data. As it appears there is no strong relationship, and the average value is skewing the data, the bmi predictor must be dropped from the data before moving onto the model building.



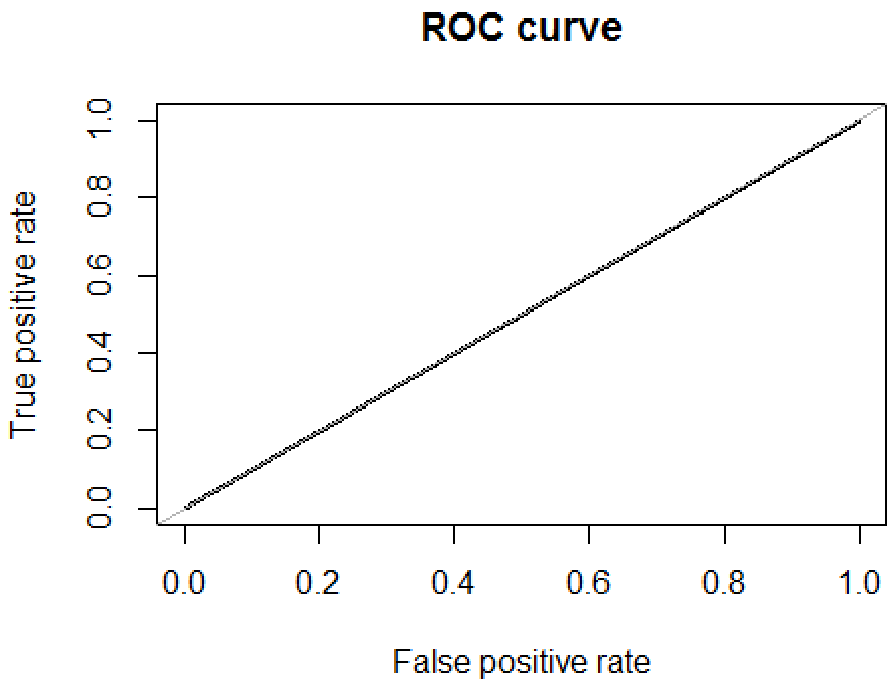
So far there is no single obvious predictor for identifying stokes. Instead, each predictor shows some weak predictive power. Also as we have seen the data is skewed with very little data for stroke cases.

Before building the first model it is necessary to to remove the bmi and id predictors. The NA values will cause errors in the machine learning algorithms and the ids shouldn't have any predictive power. A training set and test set are then partitioned from the data, again at 9 to 1 ratio due to the size of the data.

The first model, a decision tree, is trained and tested giving the results below:

Decision Tree Confusion Matrix

	1	0
1	0	0
0	13	243



```
## Area under the curve (AUC): 0.500

## [1] NA
```

It is obvious that the unbalanced nature of the data is effecting the model. To remedy this, the ROSE package will be used to try several approaches of data balancing. Over-sampling replicates random observations in the minority class until meeting a predefined ratio, this approach has a greater risk of overfitting the data. Under-sampling is a method of randomly removing observations in the majority class until meeting a predefined ratio, this approach loses a significant portion of the data. The ROSE package makes implementing both of these methods easy, and provides two additional methods. The first is a combination of the over-sampling and under-sampling methods, and the second method, refered to as 'Rose', creates synthetic data that is inserted into the minority class. The resulting dataset sizes and the ROC curves when using these methods is shown below.

Data Balancing Results

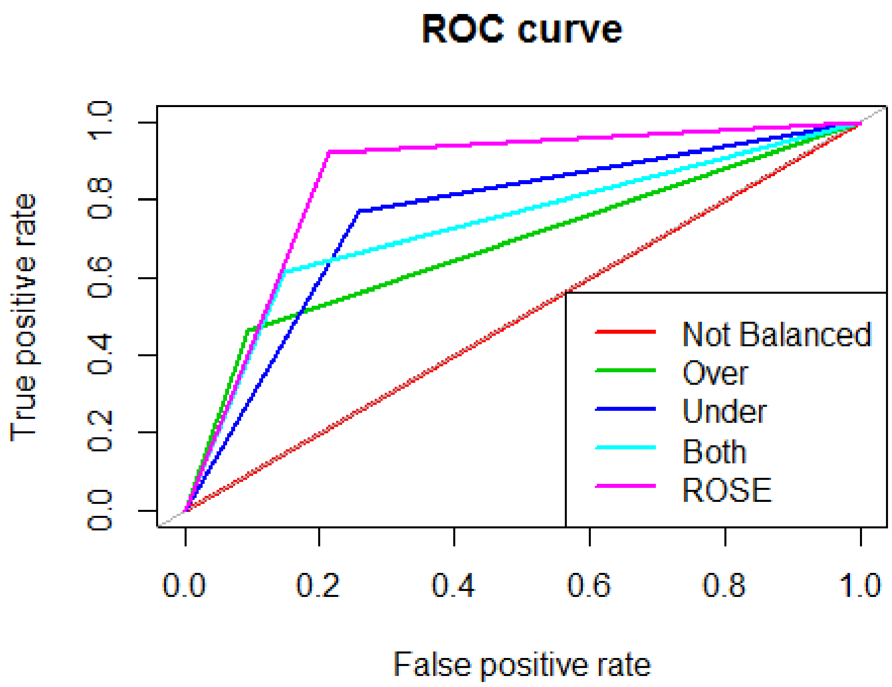
	1	0
Unbalanced	116	2183
Over-Balanced	2183	2183
Under-Balanced	116	116
Both	1136	1163
ROSE (Data Injection)	1181	1118

```
## Area under the curve (AUC): 0.500
## Area under the curve (AUC): 0.683

## Area under the curve (AUC): 0.755

## Area under the curve (AUC): 0.734

## Area under the curve (AUC): 0.855
```



The ROSE method of injecting synthetic data gives the best results. So this data is used with several other machine learning algorithms, the results of which are shown below.

Decision Tree Confusion Matrix

	1	0
1	12	52
0	1	191

```
## Area under the curve (AUC): 0.855
```

Generalized Linear Model Confusion Matrix

	1	0
1	12	52
0	1	191

```
## Area under the curve (AUC): 0.855
```

K Nearest Neighbors Confusion Matrix

	1	0
1	11	58
0	2	185

```
## Area under the curve (AUC): 0.804
```

Random Forest Confusion Matrix

	1	0
1	12	60
0	1	183

```
## Area under the curve (AUC): 0.838
```

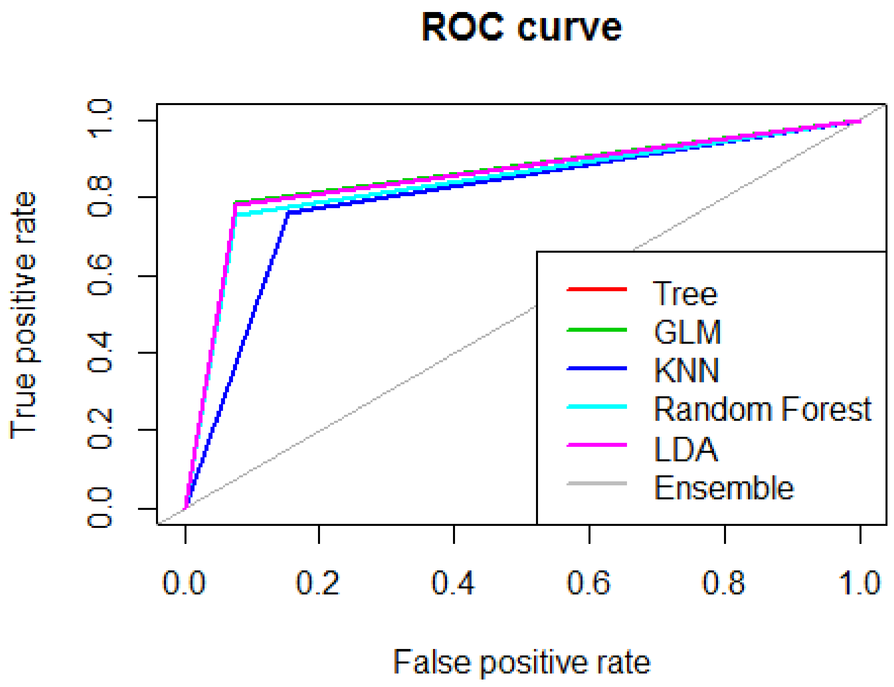
Linear Discriminate Analysis Confusion Matrix

	1	0
1	12	53
0	1	190

```
## Area under the curve (AUC): 0.852
```

Ensemble Confusion Matrix

	1	0
1	12	56
0	1	187



F1 Values

Method	F
Tree	0.3116883
GLM	0.3116883
KNN	0.2682927
Random Forest	0.2823529
LDA	0.3076923
Ensemble	0.3076923

The algorithms used were the decision tree, a generalized linear model, K nearest neighbors, a random forest, linear discriminate analysis, and an ensemble which combined all of these methods. The 'GLM' and 'Tree' methods tied for the best F1 score. The 'GLM' model was chosen to train the final model before testing on the validation set. First all the 'ROSE' method was used to inject data into the new training set. The final model was trained with this data. The validation set was munged to be in the correct format, and used to generate predictions.

Results

The final model results are shown below.

Validation set Confusion Matrix

	1	0
1	82	623
0	38	1811

```
## [1] "F value:  0.198787878787879"
```

Conclusions

This project took in multiple predictors from a confidential source and tried to predict if a patient would have a stroke. Each of the predictors, when visualized, showed some weak correlation. The final F value shows that the goal of this project can be considered inconclusive. Although the WHO accounts for it being the second leading cause of death, 11% is a difficult fraction to predict. Simply predicting 'unlikely' for every patient will yeild a high accuracy. It is worth considering that some patients that were false positive may have died from a different cause only days before a stroke occured. In addition the dataset is relatively small and NA values in predictors such as BMi only make it smaller. There may be more significant predictors that need to be found before these models become viable.

Future work with these models would include getting patient history over multiple doctor checkups. Converting each of the predictors in this data set to a time-series, would be easily possible if given access to the patients medical records and could vastly improve the results. Several other predictors such as patient ethnicity and medications would also have significant predictive power.

Citation

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset