BUAN 6356 BUSINESS ANALYTICS WITH R

Dr. Sourav Chatterjee

STROKE PREDICTION IN R

August 2, 2021

Group 4:

Dhruv Tuteja

Lukeshini Malikireddy

Maithili Dupte

Martin Vincent

Shailja Shah

TABLE OF CONTENT

SUMMARY

WHO has suggested that strokes are the second leading cause of death and third leading cause of disability worldwide, and that early detection and prevention techniques hold utmost importance. An algorithm that is able to accurately detect early warning signs of higher risk individuals could save the lives of millions. This algorithm could be useful in the health insurance industry, in which a greater understanding of the policyholder's health can maximize profitability for the company and fairness for the purchaser. Assuming it can predict strokes with relative accuracy, the company will charge a higher premium to the individuals with higher risk, so that it is able to make a profit on them. The algorithm is meant to be used by the underwriting team at an insurance company, in which it will enhance business intelligence and policy writing abilities.

The dataset was sourced from Kaggle. It was further explored and cleaned in order to deal with the null values and insignificant variables. As it was unbalanced, it had to be balanced using undersampling and oversampling. It was then split into a test and a training dataset.

Next, exploratory data analysis was carried out. Correlation between numerical variables was calculated, and bar graphs and histograms were plotted.

Random Forest, Decision Tree, and Logistic Regression models were built using normal training data, oversampling training data, undersampling training data, and both over and undersampling training data.

For our model, the sensitivity/recall is of more significance than the specificity, as it is more important to catch the true positives. Taking this into account, the best models are the random forest model built using the undersampled data, the decision trees model built using undersampled and oversampled data, and the logistic regression model built using undersampled and oversampled data.

INTRODUCTION

Stroke is when the supply of blood is prevented or partly interrupted to the brain causing a hindrance in the supply of oxygen and nutrients. Due to this, the brain cells known as the neurons begin to die. Neurons are cells which lack the capacity to renew, regenerate, and divide, causing permanent disability in stroke survivors. The World Health Organization has suggested that strokes are the second leading cause of death and third leading cause of disability worldwide. Scientists and analysts are researching early detection and prevention of strokes to reduce the severity or prevent cases altogether.

The most important consideration when building this algorithm is to ask the question, “how can this be applied to a business scenario to improve profitability for a company?” The answer is that this algorithm will be marketed towards insurance companies. In the insurance industry, decisions that will lead to profitable or non-profitable outcomes are very speculative, so any amount of improved intelligence could make for a much bigger profit margin.

In the American Insurance Industry, underwriters, who work for the insurance company, will write different policies with different premiums for each individual purchaser. There can be a high variability in the cost of these policies for seemingly similar insurance purchasers. This is because underwriters typically use black box algorithms to create policies, which can be intricate and difficult for people to understand. The obvious downside to these algorithms is that it can be hard to understand the “why” behind many of their decisions, and the upside is that they typically predict much better than humans.

The purpose of this algorithm is to predict who will have a stroke. With strokes being the second leading cause of death and third leading cause of disability, being able to predict their occurrence more accurately will have a large overall impact on the accuracy of the insurance company’s intelligence. This model is meant to be used in conjunction with other algorithms that can predict diseases, and algorithms that can determine how these predictions can be used to make profitable policies, however, that is beyond the scope of this project. This will be an important puzzle piece in the bundle of algorithms an insurance company uses to make better predictions, and ultimately, lead to more profitable policies being issued by the company.

Our goal is to build a machine learning algorithm that can accurately detect early warning signs of higher risk individuals, which could save the lives of millions. “Strokes can only occur to the elderly” and “strokes are not preventable or treatable” are common myths that persist. Another widely spread myth is that strokes are not hereditary, when in reality, strokes do run in the family. Our model aims to put aside myths and use data to predict who is likely to have a stroke.

SUGGESTED HYPOTHESIS

- Null Hypothesis (H0): There is no relationship between X (predictor variables), which includes gender, age, hypertension, heart disease, average glucose level, BMI, gender, marital status, work type, residence type, and smoking status, and Y (the response variable), which is the occurrence of stroke.

- Alternate Hypothesis (H1): There is a relationship between X (predictor variables), which includes gender, age, hypertension, heart disease, average glucose level, BMI, gender, marital status, work type, residence type, and smoking status, and Y (the response variable) which is the occurrence of stroke.

INTRODUCTION TO THE DATASET

The dataset sourced from Kaggle consists of 5110 observations in total. The columns included are ID (numeric), gender (categorical), age (numeric), hypertension (binary), heart disease (binary), ever married (categorical), work type (categorical), residence type (categorical), average glucose level (numeric), body mass index (numeric), smoking status (categorical), and ever suffered a stroke (binary).

The entire data is summarized as below:

GENDER

Male	2115
Female	2994
Other	1

HYPERTENSION

Hypertension	4612
No hypertension	498

MARITAL STATUS

Never married	1757
Been married	4898

OCCUPATION

children	687
Government employee	657
Never worked	22
Private sector employee	2925
Self employed	819

RESIDENCE TYPE

Rural	2514
Urban	2596

SMOKING STATUS

Formerly smoked	885
Never smoked	1892
Smokes presently	789
unknown	1544

PREVALANCE OF HEART DISEASE

Suffers heart disease	276
No heart disease	4834

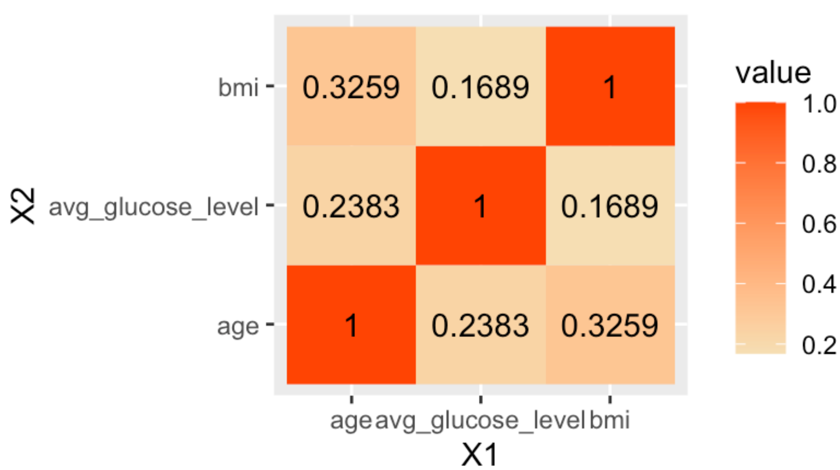
DATA SUMMARY

VARIABLE	FREQUENCY	VARIABLE	FREQUENCY
id	Min. : 67	work_type	Length:5110
	1st Qu.:17741		Class: character
	Median :36932		Mode: character
	Mean :36518	Residence_type	Length:5110
	3rd Qu.:54682		Class: character
	Max. :72940		Mode: character
gender	Length:5110	avg_glucose_level	Min.: 55.12
	Class: character		1st Qu.: 77.25
	Mode: character		Median: 91.89
age	Min.: 0.08		Mean :106.15
	1st Qu.:25.00		3rd Qu.:114.09
	Median :45.00		Max. :271.74
	Mean :43.23	bmi	Length:5110
	3rd Qu.:61.00		Class: character
	Max. :82.00		Mode: character
hypertension	Min. :0.00000	smoking_status	Length:5110
	1st Qu.:0.00000		Class: character
	Median :0.00000		Mode: character
	Mean :0.09746	stroke	Min. :0.00000
	3rd Qu.:0.00000		1st Qu.:0.00000
	Max. :1.00000		Median :0.00000
heart_disease	Min. :0.00000		Mean :0.04873
	1st Qu.:0.00000		3rd Qu.:0.00000
	Median :0.00000		Max. :1.00000
	Mean :0.05401		
	3rd Qu.:0.00000		
	Max. :1.00000		
ever_married	Length:5110		
	Class: character		
	Mode: character		

DATA CLEANING

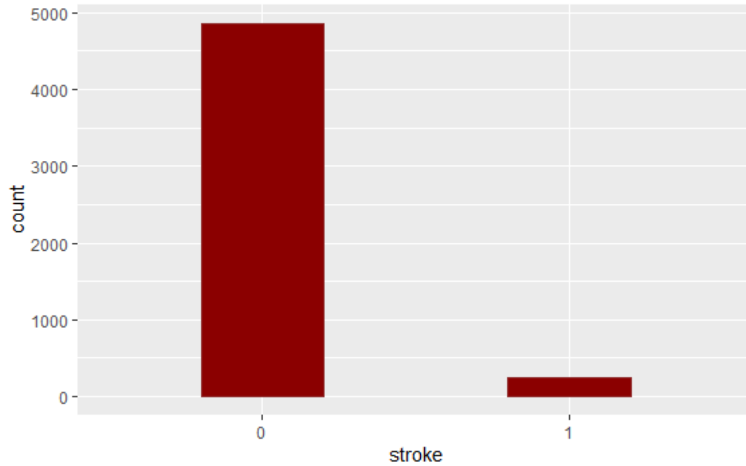
- The selected data set was imported into R-studio and first checked for the prevalence of null values. The variable BMI showed the presence of null values. This was replaced with the mean BMI values.
- The N/A values in the BMI column were converted to NA.
- The gender column had only one entry as ‘other’, and this row was removed.
- The ID column held no significance in the prediction model, so it was dropped.
- Following this, the categorical columns were converted to factors. The dataset was split into a training dataset and a test dataset. The training dataset will be used in building the prediction model, and the test dataset will be used for examining the performance of the model.
- We discovered that in the dataset the people who had suffered a stroke outnumbered those who had not, which made the dataset highly imbalanced. To solve this, the training dataset was balanced by undersampling and oversampling techniques using the ROSE library.

EXPLORATORY DATA ANALYSIS



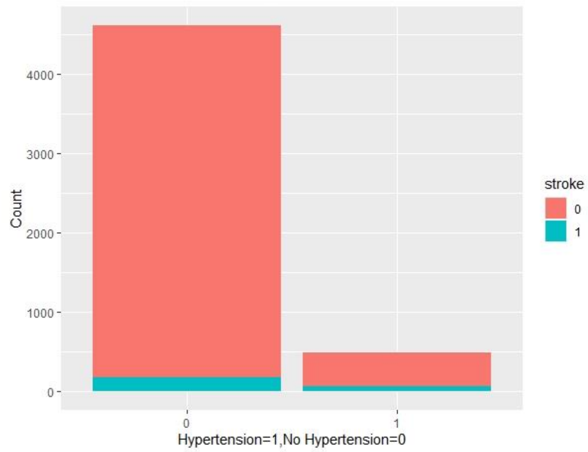
Correlation Between the Numeric Variables

The above table shows the correlation between the numeric variables – age, average glucose level, and BMI. All of the numeric variables are positively correlated to each other, and the highest correlation is in between average glucose level and BMI is 0.3259.



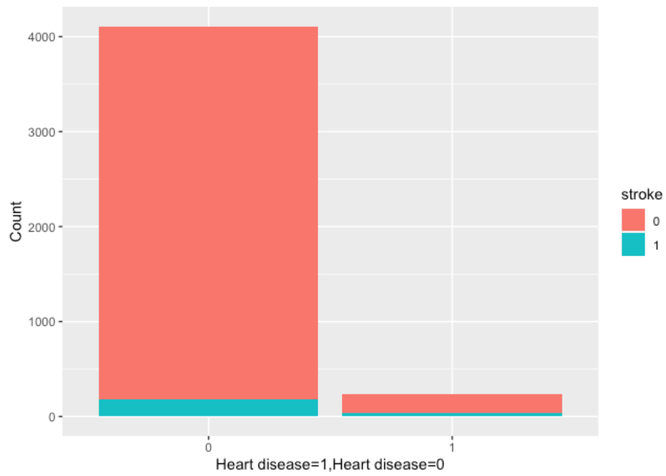
Frequency of Strokes in the Dataset

The bar graph above showcases the number of people who have suffered a stroke and the number of people who have not. Out of 5110 observations, 249 had a stroke and 4861 did not. On the X-axis 0 denotes non-stroke and 1 denotes stroke.



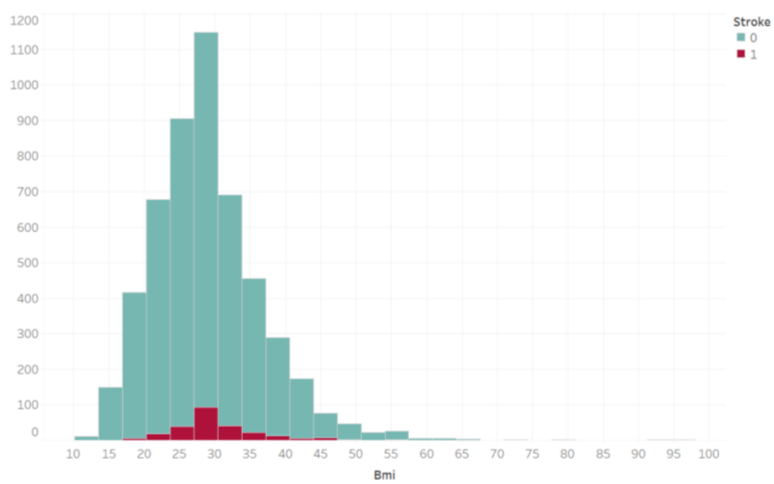
Frequency of Hypertension Grouped by Stroke

The bar graph above denotes the number of observations with and without hypertension (0 and 1) who have suffered a stroke (red and blue colors). The graph suggests that people with hypertension have a higher chance of suffering a stroke.



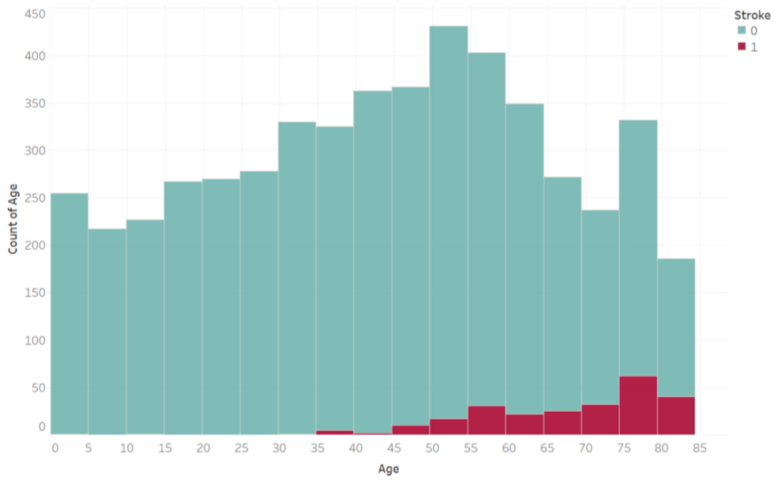
Frequency of Heart Disease Grouped by Stroke

The bar graph above shows the number of people with and without heart disease (0 and 1) who have suffered a stroke (red and blue colors).



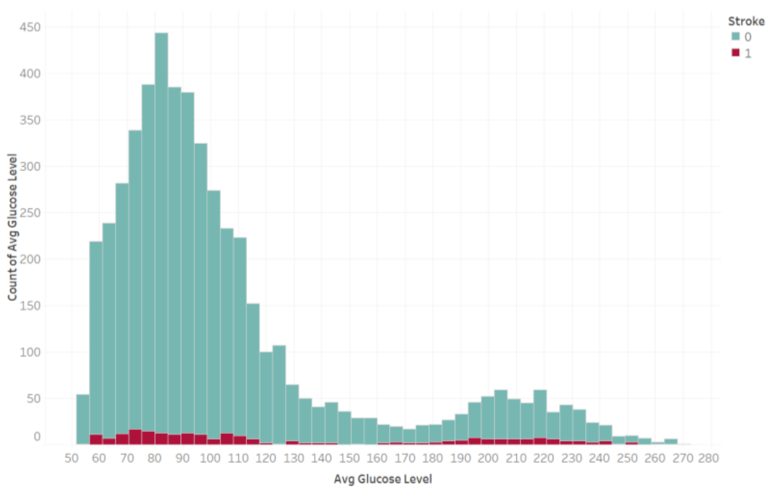
BMI Histogram

The histogram above demonstrates the number of people whose BMI falls within a particular range, sorted by those who have and have not suffered a stroke. The highest observation frequency and stroke frequency lie between the 27.5 to 30 BMI range.



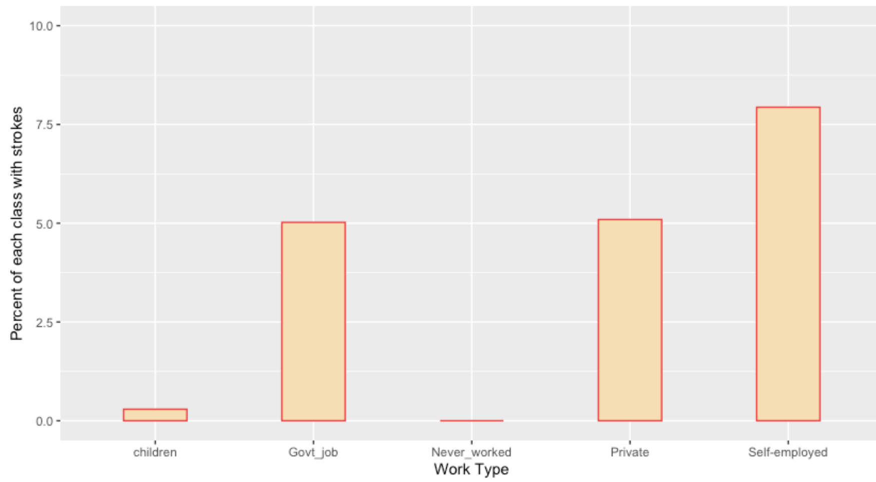
Age Histogram

The histogram above displays the age distribution based on stroke and non-stroke occurences. According to the graph, most of the people for any age group have not suffered a stroke. Among those who have suffered a stroke, the highest frequency range is 75-80 years old.



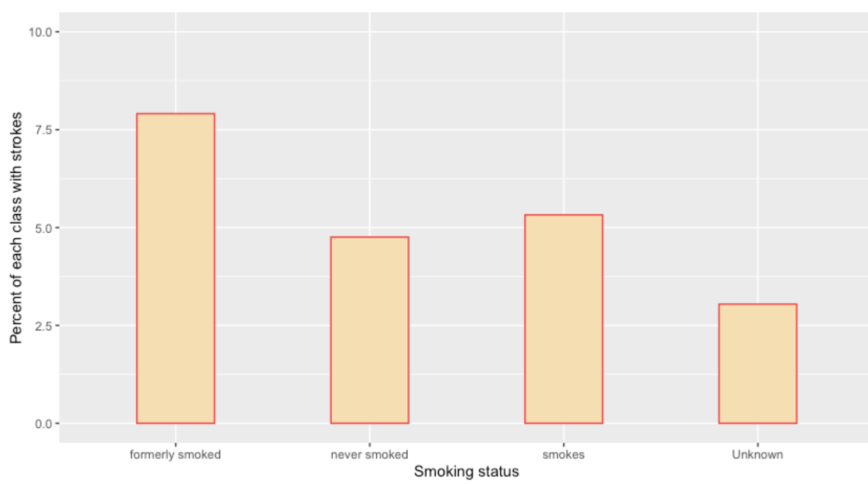
Average Glucose Level Histogram

The histogram above displays the average glucose level distribution based on stroke and non-stroke occurences.



Percentage of Strokes Grouped by Work Type

The bar graph above shows the percentage of strokes based on different working types. Self employed individuals have the highest rate, followed by individuals with private and government jobs.



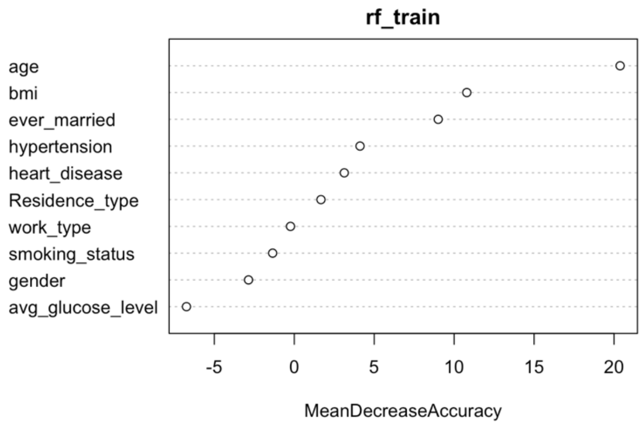
Percentage of Stroke Grouped by Smoking Status

The bar graph above shows the rate of strokes by smoking status. Based on the graph, those who formerly smoked have the highest rate, followed by smokers. It is expected that smokers would have the highest rate, but that is not true in this dataset.

PREDICTIVE MODEL BUILDING

- **RANDOM FOREST MODELS**

1. USING THE NORMAL TRAINING DATA

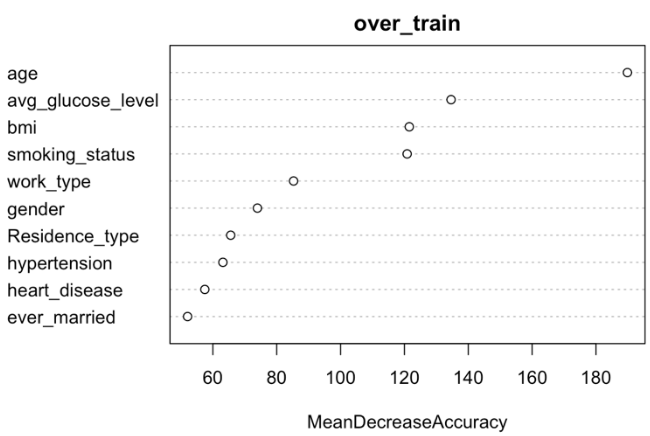



```
          Reference
Prediction   0   1
         0 971  48
         1   2   0

               Accuracy : 0.951
                 95% CI : (0.9359, 0.9634)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 0.6513

                  Kappa : -0.0038

 Mcnemar's Test P-Value : 1.966e-10

            Sensitivity : 0.000000
            Specificity : 0.997945
         Pos Pred Value : 0.000000
         Neg Pred Value : 0.952895
             Prevalence : 0.047013
         Detection Rate : 0.000000
   Detection Prevalence : 0.001959
      Balanced Accuracy : 0.498972

       'Positive' Class : 1
```

Variable importance plot for Random Forest using the normal training data

The model built using the normal training data produced the results shown above. The accuracy of this model was 95.1% and the sensitivity was 0.000000.

2. USING THE OVERSAMPLED TRAINING DATA

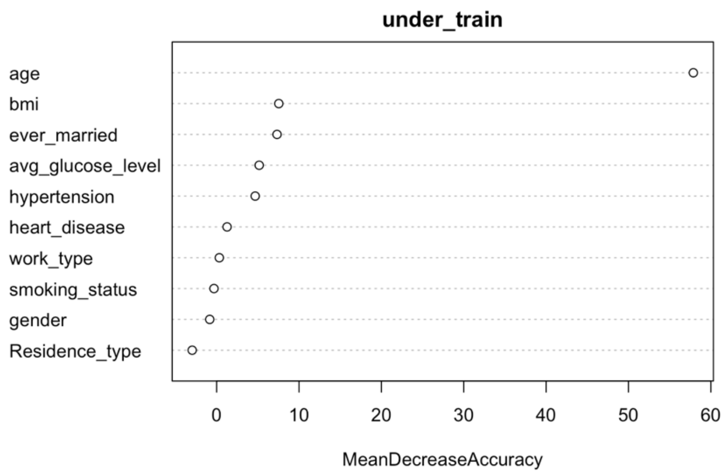



```
          Reference
Prediction   0   1
         0 959  42
         1  14   6

               Accuracy : 0.9452
                 95% CI : (0.9294, 0.9583)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 0.8935445

                  Kappa : 0.153

 Mcnemar's Test P-Value : 0.0003085

            Sensitivity : 0.125000
            Specificity : 0.985612
         Pos Pred Value : 0.300000
         Neg Pred Value : 0.958042
             Prevalence : 0.047013
         Detection Rate : 0.005877
   Detection Prevalence : 0.019589
      Balanced Accuracy : 0.555306

       'Positive' Class : 1
```

Variable importance plot for Random Forest using the oversampled data

The model built using the oversampled training data produced the results shown above. The accuracy of this model was 94.5% and the sensitivity was 0.125000.

3. USING THE UNDERSAMPLED TRAINING DATA

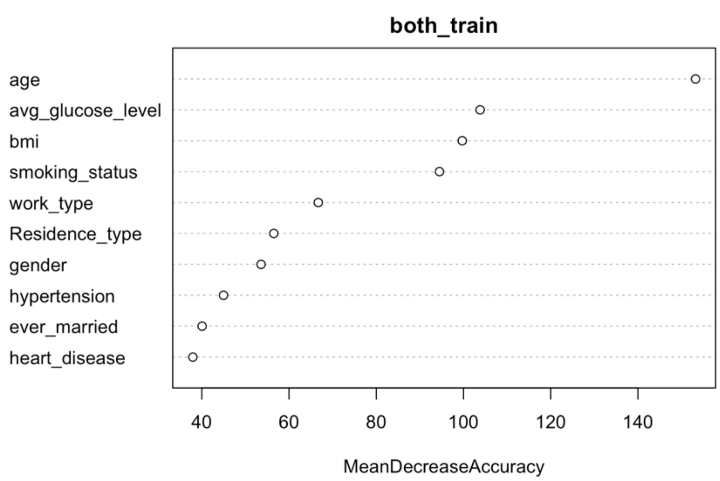



```
          Reference
Prediction   0   1
         0 732  15
         1 241  33

               Accuracy : 0.7493
                 95% CI : (0.7215, 0.7756)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1358

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.68750
            Specificity : 0.75231
         Pos Pred Value : 0.12044
         Neg Pred Value : 0.97992
             Prevalence : 0.04701
         Detection Rate : 0.03232
   Detection Prevalence : 0.26836
      Balanced Accuracy : 0.71991

       'Positive' Class : 1
```

Variable importance plot for Random Forest using the undersampled data

The model built using the undersampled training data produced the results shown above. The accuracy of this model was 74.93% and the sensitivity was 0.68750.

4. USING BOTH UNDER AND OVERSAMPLED DATA



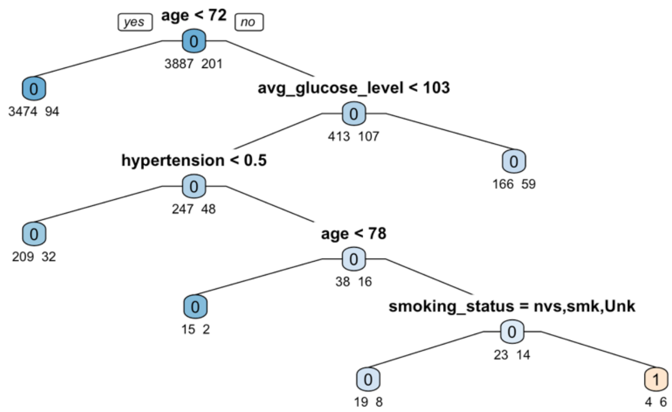

```
          Reference
Prediction   0   1
         0 892  30
         1  81  18

               Accuracy : 0.8913
                 95% CI : (0.8706, 0.9097)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1938

 Mcnemar's Test P-Value : 2.077e-06

            Sensitivity : 0.37500
            Specificity : 0.91675
         Pos Pred Value : 0.18182
         Neg Pred Value : 0.96746
             Prevalence : 0.04701
         Detection Rate : 0.01763
   Detection Prevalence : 0.09696
      Balanced Accuracy : 0.64588

       'Positive' Class : 1
```

Variable importance plot for Random Forest using under and oversampled training data

The model built using the under and oversampled training data produced the results shown above. The accuracy of this model was 89.13% and the sensitivity was 0.37500.

- **DECISION TREE MODELS**

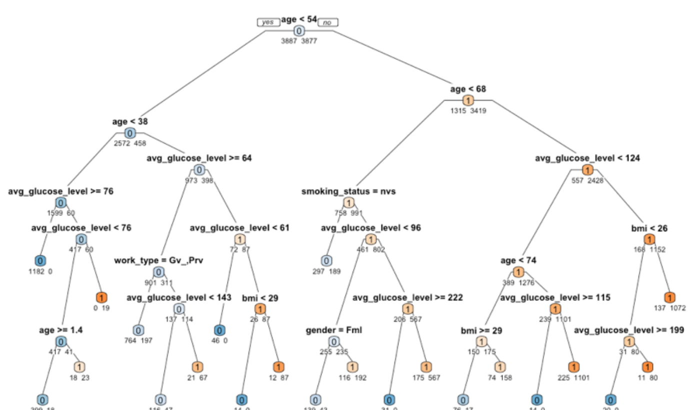
1. USING THE NORMAL TRAINING DATA




```
          Reference
Prediction   0   1
         0 971  48
         1   2   0

               Accuracy : 0.951
                 95% CI : (0.9359, 0.9634)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 0.6513

                  Kappa : -0.0038

 Mcnemar's Test P-Value : 1.966e-10

            Sensitivity : 0.000000
            Specificity : 0.997945
         Pos Pred Value : 0.000000
         Neg Pred Value : 0.952895
             Prevalence : 0.047013
         Detection Rate : 0.000000
   Detection Prevalence : 0.001959
      Balanced Accuracy : 0.498972

       'Positive' Class : 1
```

The model built using the normal training data produced the results shown above. The accuracy of this model was 95.1% and the sensitivity was 0.000000.

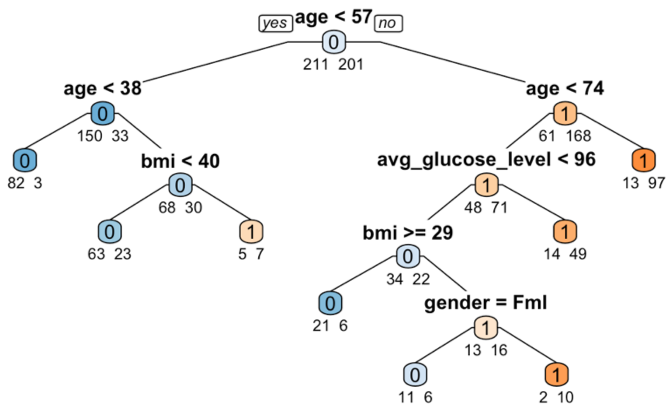
2. USING THE OVERSAMPLED DATA




```
          Reference
Prediction   0   1
         0 754  17
         1 219  31

               Accuracy : 0.7689
                 95% CI : (0.7418, 0.7944)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1402

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.64583
            Specificity : 0.77492
         Pos Pred Value : 0.12400
         Neg Pred Value : 0.97795
             Prevalence : 0.04701
         Detection Rate : 0.03036
   Detection Prevalence : 0.24486
      Balanced Accuracy : 0.71038

       'Positive' Class : 1
```

The model built using th oversampled training data produced the results shown above. The accuracy of this model was 76.89% and the sensitivity was 0.64583.

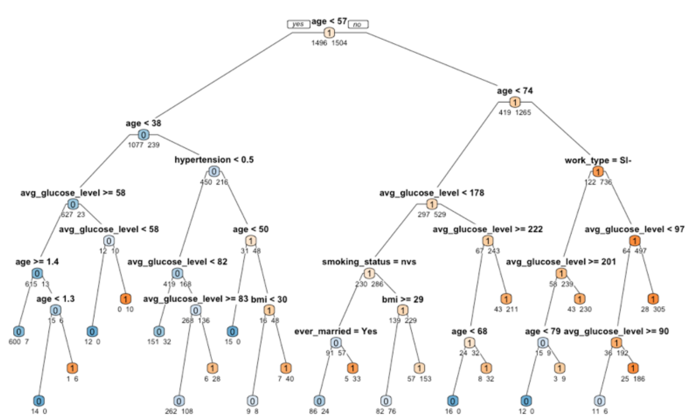
3. USING THE UNDERSAMPLED DATA




```
          Reference
Prediction   0   1
         0 714  17
         1 259  31

               Accuracy : 0.7297
                 95% CI : (0.7013, 0.7567)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1118

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.64583
            Specificity : 0.73381
         Pos Pred Value : 0.10690
         Neg Pred Value : 0.97674
             Prevalence : 0.04701
         Detection Rate : 0.03036
   Detection Prevalence : 0.28404
      Balanced Accuracy : 0.68982

       'Positive' Class : 1
```

The model built using the undersampled training data produced the results shown above. The accuracy of this model was 72.97% and the sensitivity was 0.64583.

4. USING BOTH OVER AND UNDERSAMPLED DATA

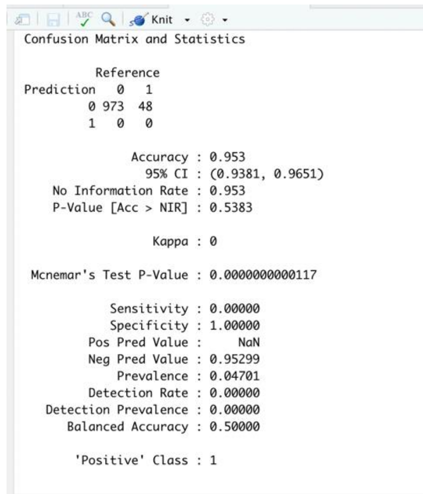



```
          Reference
Prediction   0   1
         0 787  18
         1 186  30

               Accuracy : 0.8002
                 95% CI : (0.7743, 0.8243)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1629

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.62500
            Specificity : 0.80884
         Pos Pred Value : 0.13889
         Neg Pred Value : 0.97764
             Prevalence : 0.04701
         Detection Rate : 0.02938
   Detection Prevalence : 0.21156
      Balanced Accuracy : 0.71692

       'Positive' Class : 1
```

The model built using the over and undersampled training data produced the results shown above. The accuracy of this model was 80.02% and the sensitivity was 0.62500.

- **LOGISTIC REGRESSION MODEL**

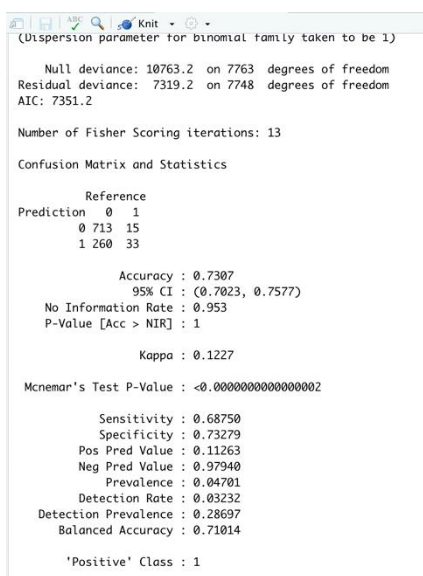
1. USING THE NORMAL TRAINING DATA



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 973  48
         1   0   0

               Accuracy : 0.953
                 95% CI : (0.9381, 0.9651)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 0.5383

                  Kappa : 0

 Mcnemar's Test P-Value : 0.0000000000117

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.95299
             Prevalence : 0.04701
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 1
```

The logistic regression model built using normal training data produced an accuracy of 95.3% and sensitivity of 0.00000.

2. USING THE OVERSAMPLED DATA



```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10763.2  on 7763  degrees of freedom
Residual deviance:  7319.2  on 7748  degrees of freedom
AIC: 7351.2

Number of Fisher Scoring iterations: 13

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 713  15
         1 260  33

               Accuracy : 0.7307
                 95% CI : (0.7023, 0.7577)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1227

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.68750
            Specificity : 0.73279
         Pos Pred Value : 0.11263
         Neg Pred Value : 0.97940
             Prevalence : 0.04701
         Detection Rate : 0.03232
   Detection Prevalence : 0.28697
      Balanced Accuracy : 0.71014

       'Positive' Class : 1
```

The logistic regression model built using oversampled data produced an accuracy of 73.07% and sensitivity of 0.68750.

3. USING THE UNDERSAMPLED DATA

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 707  13
         1 266  35

               Accuracy : 0.7267
                 95% CI : (0.6983, 0.7539)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.13

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.72917
            Specificity : 0.72662
         Pos Pred Value : 0.11628
         Neg Pred Value : 0.98194
             Prevalence : 0.04701
         Detection Rate : 0.03428
   Detection Prevalence : 0.29481
      Balanced Accuracy : 0.72789

       'Positive' Class : 1
```

The logistic regression model built using undersampled data produced an accuracy of 72.67% and sensitivity of 0.72917.

4. USING BOTH OVER AND UNDERSAMPLED DATA

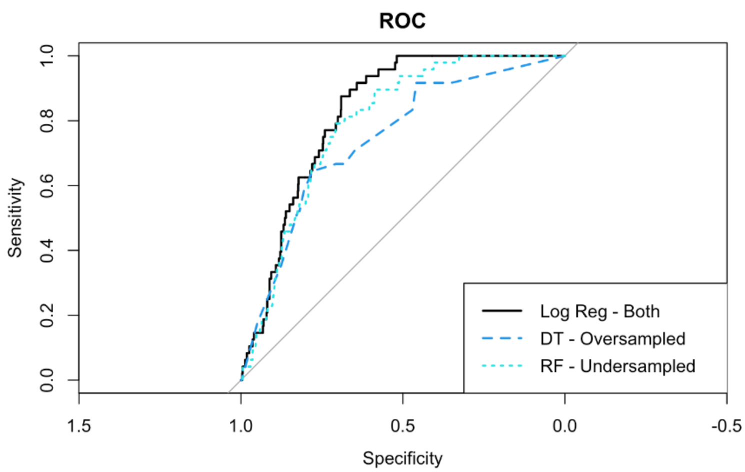
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 708  11
         1 265  37

               Accuracy : 0.7297
                 95% CI : (0.7013, 0.7567)
    No Information Rate : 0.953
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1418

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.77083
            Specificity : 0.72765
         Pos Pred Value : 0.12252
         Neg Pred Value : 0.98470
             Prevalence : 0.04701
         Detection Rate : 0.03624
   Detection Prevalence : 0.29579
      Balanced Accuracy : 0.74924

       'Positive' Class : 1
```

The logistic regression model built using both over and undersampled data produced an accuracy of 72.97% and sensitivity of 0.77083.

COMPARING THE MODELS

The class of interest is the positive diagnosis of the occurrence of stroke. Sensitivity, or recall, is the best measure of the accuracy of that. Therefore, while building the stroke prediction model, sensitivity plays a more important role. It is crucial to diagnose people who are showing higher chances of suffering a stroke, and therefore, accurately predicting the true positive value holds higher importance.

If we compare the random forest models built above, the model built using the undersampled data has the highest sensitivity of 0.68750. Amongst the decision tree models built, we got two best models - the model built using undersampling training data and the model built using oversampling training data, both of which gave a sensitivity of 0.64583. Of the four logistic regression models built, the one using both the undersampled and oversampled data had the highest sensitivity/recall of 0.77083.



ROC Curve for Best Performing Models

The ROC curve above compares the best random forest, decision tree, and logistic regression models. It shows that the best model based on sensitivity is the logistic regression model, followed by random forest, followed by decision tree. Since the logistic regression which used over and undersampling had the highest recall of 0.77083, we concluded that this is the model we should choose.

REFERENCES

1. Chakure, A. (2019). *Decision Tree Classification.* https://medium.com/swlh/decision-tree-classification-de64fc4d5aac.
2. Department of Neurology, B. U.-H. (2010). *Influence of Age and Health Behaviors on Stroke Risk: Lessons from Longitudinal Studies.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3006180/.
3. *How AI is able to Predict and Detect a Stroke* . (n.d.). https://getreferralmd.com/2019/10/how-ai-is-able-to-predict-and-detect-a-stroke/.
4. Lin, S. (2021). *Stroke Prediction Constructing prediction model for the risk of stroke.* https://medium.com/geekculture/stroke-prediction-d26c15f9d1.
5. Plapinger, T. (2017). *What is a Decision Tree?* https://towardsdatascience.com/what-is-a-decision-tree-22975f00f3e1.
6. *Stroke: Online Calculator Can Predict Your Risk* . (2020). Virgnia: https://newsroom.uvahealth.com/2020/08/13/stroke-online-calculator-can-predict-your-risk/.
7. Tony Yiu Tony Yiu Tony Yiu Data Science @Solovis, D. m.-b.-b. (2019). *Understanding Random Forest How the Algorithm Works and Why it Is So Effective.* https://towardsdatascience.com/understanding-random-forest-58381e0602d2.
8. Xuenan Peng, 1. J. (2020). *Longitudinal Average Glucose Levels and Variance and Risk of Stroke: A Chinese Cohort Study*. https://www.hindawi.com/journals/ijhy/2020/8953058/.
9. Yiu, T. (2019). *Understanding Random Forest How the Algorithm Works and Why it Is So Effective* . https://towardsdatascience.com/understanding-random-forest-58381e0602d2.