Ben Lortie, Vincent Bianchi, Kamryn Chavez, Zoe Elders

Machine Learning

2 December 2023

Machine Learning Model - Stroke Prediction

Introduction

Our group has decided to delve deeper into the key factors linked to the occurrence of a stroke. The CDC defines a *stroke* as an occurrence "when something blocks blood supply to part of the brain or when a blood vessel in the brain bursts" (About Stroke, 2023). As a result, many people face severe and sometimes life-threatening side effects caused by parts of the brain being damaged and dying. As the fifth leading cause of death in the United States, this problem has affected millions of people. We plan to create a model that will show the most significant and recurring factors in those who have experienced a stroke. This problem is interesting because if there are multiple variables, we can correlate those who have experienced a stroke in the past, identify whether they are at a high risk of suffering one, and take preventative actions to safeguard their health and well-being, thus potentially saving lives.

Throughout this project, we will research previous attempts at this problem to get a broader perspective on the pre-stroke lifestyles of those who have experienced one. Then, we will use multiple models to predict which factors have the most significant relation to the occurrence of a stroke. The models used are Logistic Regression, Lasso Regression, Random Forest, Confusion Matrix, and XGBoost. While interpreting the results, we will compile the most prominent attributes. Furthermore, we will give context regarding our chosen dataset and the methodology behind our approach to the problem. Afterward, we will go into depth on the results and findings from our models and any challenges we faced during our trials. Lastly, we will discuss the future steps we can take with the results of our data and provide valuable insights.

Related Work

This problem has been attempted before as there are many medical research studies regarding the prevention of strokes and pre-stroke lifestyle, as stated in the academic journal "Pre-stroke physical activity and admission stroke severity: A systematic review" (Hung SH, Ebaid D, Kramer S, et a, 2021). This study emphasizes the lifestyle type's role in previous stroke experiences. It concluded that there was a significant relationship between physical activity and

stroke severity based on the observational studies. Another study that took another approach when looking into the predictors of strokes was "Early Stroke Prediction Methods for Prevention of Strokes" (Kaur et al., 2022). This study focused on the bioelectrical signals to forecast strokes and help save a life if the stroke risk appears during an early stage. Our approach differs from much of the research we found as most focus on biological predictors such as nerves and muscles. We are taking a broader look at lifestyle choices such as smoking, marriage, and work type.

Data Description

The group utilized a publicly available dataset from Kaggle containing relevant patient health information. This dataset included a specific column indicating whether each individual had experienced a stroke. There were 4908 samples and 11 variables (excluding the irrelevant 'id' column). The particular variables included in the dataset were gender, age, hypertension (1 or 0), history of heart disease (1 or 0), if a patient was ever married (Yes or No), work type, residence type, average glucose level, BMI, smoking status, and the response variable, stroke (1 or 0). To prepare our dataset for modeling, a few preprocessing steps were necessary. One of the first steps taken in handling variable data types was treating strings and dummy characters as factors for proper data analysis. After that, we decided to omit missing data from our dataset to ensure our analysis was based only on completed data entries. Next, we used dummy coding to treat categorical variables within the dataset. Finally, we removed any rows where gender was marked as "Other" to avoid any confusion about the effect of gender on predicted stroke status within the model. The training set for our model was created by sampling 70% of the total observations, with the final 30% used for model testing. Data exploration graphs are included in Figures 1 - 7 in the appendix.

Methods

Our methodology involved the development of a set of models to discern the most optimal one within the contextual framework of the problem at hand - namely, the accurate prediction of future strokes. The four models created were logistic regression, lasso, random forest, and boosting model. The first model, logistic regression, was the easiest to create and interpret. The training of the model, along with the implementation, was straightforward. The downside to the logistic regression model is that it cannot capture complex relationships and is sensitive to noisy data. Next, we created a cross-validated lasso model to find the most

significant variables contributing to a future stroke. The variables not brought to zero after setting lambda parameters were deemed to have a statistically significant effect on the prediction of strokes. The two models that were created were the random forest and boosting models. The random forest model included a comprehensive series of decision trees built in parallel. This model was more complex and intense to train, but it produced more valuable information about the data. The boosting model is similar to the random forest, but the trees were built sequentially. One key strength of this model is that it has the most substantial predictive power.

Results

For this classification problem, we used confusion matrices to analyze the predictive power of our models. Due to the specificity-sensitivity tradeoff, we had to contemplate which metric to maximize. In the case of stroke prediction, we decided that maximizing sensitivity was more important (i.e., maximizing the proportion of actual positive cases that were predicted to be positive). Ultimately, we would want to reduce false positive and false negative counts, but focusing on keeping the false negative rate low is safer. This way, patients who are deemed to not be at risk for a stroke are at shallow risk. Due to the imbalance of the response variable, accuracy alone can not be relied upon to measure model quality, hence the evaluation of balanced accuracy. As such, we wanted to evaluate our models based on the proportional accuracy measure while simultaneously aiming to achieve a reasonably high sensitivity. As a secondary priority, we continuously increased specificity while keeping sensitivity high.

To start, we created a logistic regression model that would return an equation that defines the probability of one's stroke occurrence based on several determining factors. Using a cutoff value of 0.1, it achieved a balanced accuracy of 0.684, as seen in the confusion matrix in Figure 8. While the specificity value is high, scoring around 0.876, sensitivity only scores around 0.492. Due to the complexity of the actual variable relationships, we deemed the logistic regression model unable to adequately capture the relationship between the predictors and the response variable. As such, we looked at our output skeptically, unable to confidently say if the variables had a significant effect. To supplement, we generated a lasso regression model to understand better which variables could substantially affect stroke occurrence. This analysis, cross-validated across a sequence of lambda values, gave us a list of variables with non-zero coefficients. These variables were age, hypertension, heart disease, and average glucose level, along with certain

levels of the factor variables work type and smoking status. Moving on to our other models, we looked for these variables as we believed their significance could be validated.

After regression, we created a random forest model. While a singular decision tree would have been easier to make, it is more prone to overfitting and does not have strong predictive power relative to the random forest. In contrast, combining multiple decision trees makes the random forest model more robust. This model's prediction evaluation used the same cutoff value of 0.1. While accuracy scored slightly lower at 0.826, balanced accuracy increased to 0.695, according to the confusion matrix in Figure 9. Importantly, specificity remains very high at 0.839 while sensitivity increased, scoring at 0.552. While there was still some work to be done to find a more accurate model, we found that the random forest model was much better than logistic regression. The random forest model captured more complex relationships between the variables, leading to a higher balanced accuracy.

Our final effort was an XGBoost model, which was similar to the random forest model except that the trees were being built sequentially now instead of in parallel. A different cutoff value was used this time, as we dropped it to 0.04. Again, according to the confusion matrix in Figure 10, while the overall accuracy was lower than the previous model, this time at 0.781, balanced accuracy jumped quite a bit to 0.764 - a substantial improvement on the other models. Specificity was still high at 0.782, while sensitivity increased heavily, scoring at 0.746.

As a concluding effort, we wanted to plot the AUC values for all three prediction models, as seen in the appendix charts. According to the ROC plot in Figure 11, the XGBoost has the highest AUC at 0.83, followed closely by (interestingly) logistic regression at 0.82 and random forest at 0.79. The 2nd and 3rd place rankings come as a surprise as we had previously predicted that random forest would be ahead of logistic regression; however, XGBoost scoring first in this aligns with our predictions.

Discussion

The key findings from this model indicated that age, average glucose level, and BMI are the top three leading variables in predicting a stroke, according to Figure 12. These key findings are unsurprising since the medical community already widely accepts all these leading variables. As stated by the Stroke Awareness Foundation, obesity, diabetes, and age over 65 are listed as putting you at risk of a stroke, as we predicted with our model. These variables are also the leading causes of high blood pressure. Your risk of developing high blood pressure tends to

increase with age, especially when accompanied by conditions like diabetes and obesity. The Stroke Awareness Foundation states, “High blood pressure, also known as hypertension, can damage arteries and put you at risk for stroke as much as 4-fold”. This reassures our critical findings that age, average glucose level, and BMI are high-risk variables in predicting a stroke. That, with the combination of the current medical stroke research, makes our model a confirmatory study in leading causes of strokes.

While discussing our findings in strokes, we think it is important to note that all variables and risk factors can be due to lifestyle choices. It is common knowledge that a healthy diet and exercise can help maintain BMI, glucose levels, and blood pressure. “In the United States, about 795,000 people suffer a stroke each year. Someone has a stroke every 40 seconds, and every 4 minutes someone dies from stroke” (Stroke Awareness Foundation). Knowing those statistics on strokes and how common they are is a very frightening thing to learn, but it should cause some action in your lifestyle. As we know, older age is a risk factor. Starting now to implement good lifestyle choices will significantly decrease your chances of getting a stroke in the future. Strokes are preventable.

Conclusion and Future Work

Strokes continue to be an incredibly severe health concern in the world, and though they are largely preventable, more work can still be done to help at-risk patients before a stroke occurs. Our goal was to build a series of machine learning models - logistic regression with lasso, random forest, and XGBoost - to predict the stroke risk status of our sample patients. Our XGBoost model was the most accurate as we could correctly indicate and identify the risk status of over 75% of our sample. In terms of applying our work, we plan to use our models to larger data sets, continuously fine-tuning them to become increasingly accurate. By doing this, we can provide healthcare and medical workers with an extra resource to help their patients as they seek to deliver the required care.

Contributions

- Introduction - Zoe
- Related Work - Zoe
- Data Description - Vincent
- Methods- Vincent
- Results - Ben
- Discussion - Kamryn
- Conclusion and Future Work- Kamryn
- R Code - Ben

Bibliography

Centers for Disease Control and Prevention. (2023, May 4). *About stroke*. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/about.htm#:~:text=A%20stroke%2C%20sometimes%20called%20a,term%20disability%2C%20or%20even%20death.

Fedesoriano. (2021, January 26). *Stroke prediction dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Hung SH, Ebaid D, Kramer S, et al. Pre-stroke physical activity and admission stroke severity: A systematic review. International Journal of Stroke. 2021;16(9):1009-1018. doi:10.1177/1747493021995271

Kaur, M., Sakhare, S. R., Wanjale, K., & Akter, F. (2022, April 11). *Early stroke prediction methods for prevention of Strokes*. Behavioural Neurology. https://www.hindawi.com/journals/bn/2022/7725597/

Stroke Awareness Foundation. “Stroke Risk Factors.” *Stroke Awareness Foundation*, 31 Jan. 2021, www.strokeinfo.org/stroke-risk-factors/.

Stroke Awareness Foundation. “Stroke Facts & Statistics.” *Stroke Awareness Foundation*, 11 July 2023, www.strokeinfo.org/stroke-facts-statistics/#:~:text=In%20the%20United%20States%2C%20about,males%20and%2060%25%20in%20females.

Charts

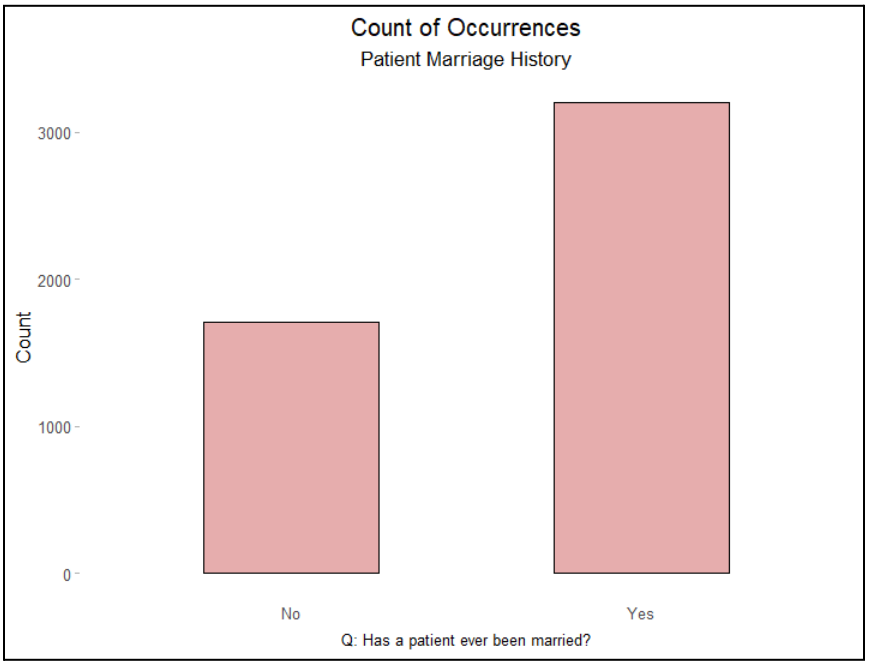


Figure 1: Column chart for the number of patients who have ever been married.

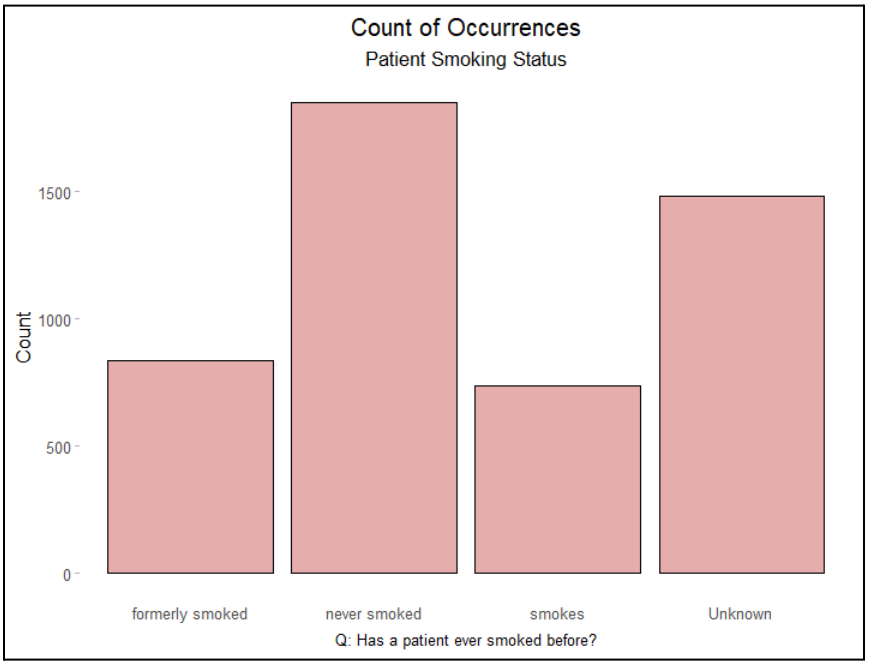


Figure 2: Column chart for the smoking status of the patients.

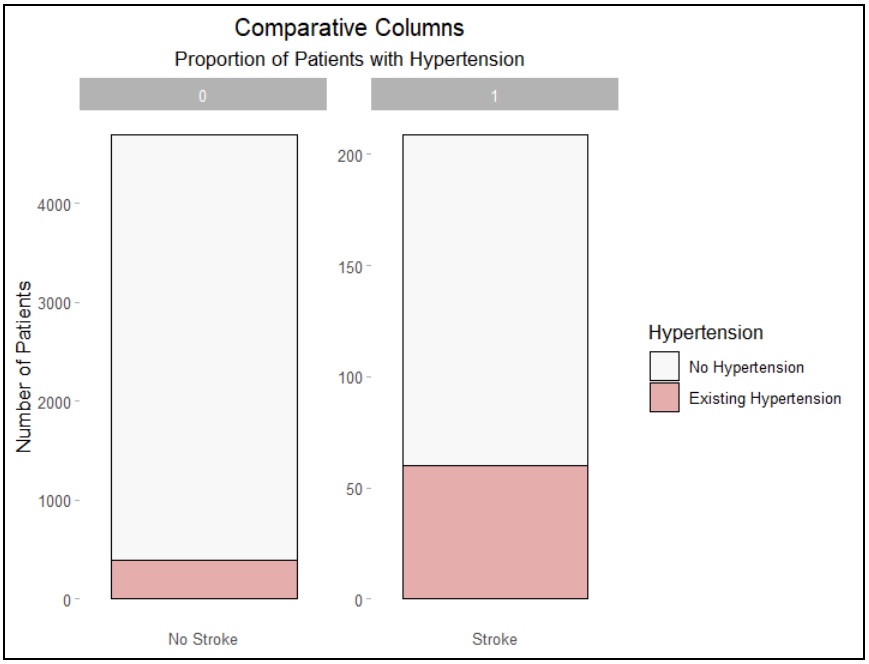


Figure 3: Stacked column chart for the proportion of patients who have hypertension, split by stroke status. The proportion of existing hypertension was higher in the sample of stroke patients than in the sample of non-stroke patients.

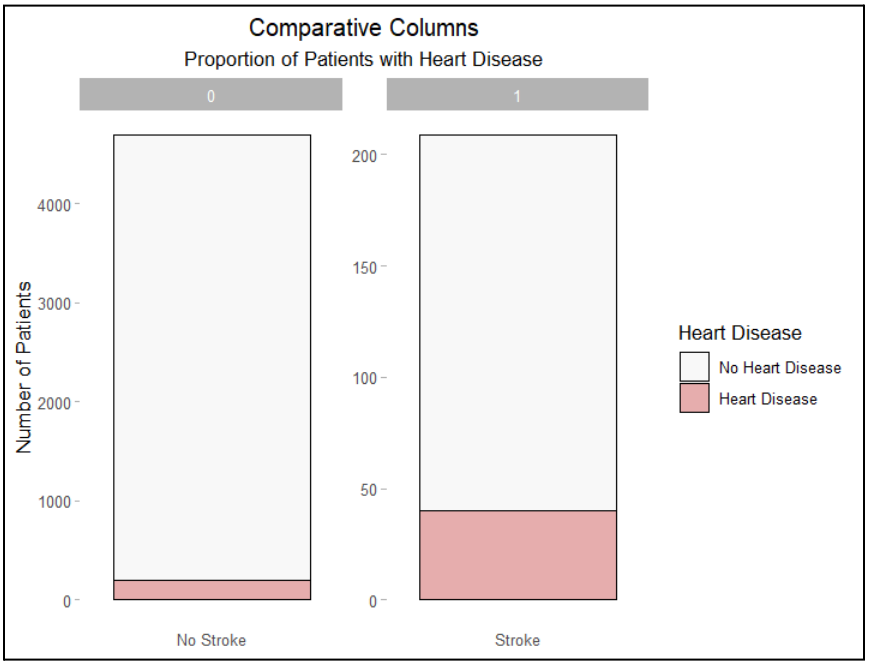


Figure 4: Stacked column chart for the proportion of patients who have heart disease, split by stroke status. The proportion of heart disease history was higher in the sample of stroke patients than in the sample of non-stroke patients.

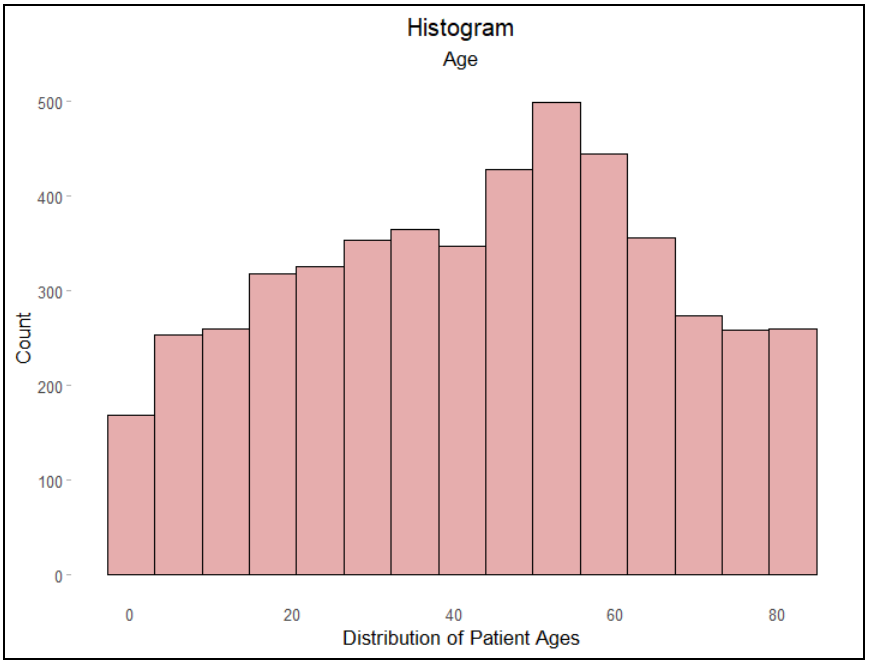


Figure 5: Histogram for the distribution of patients based on age.

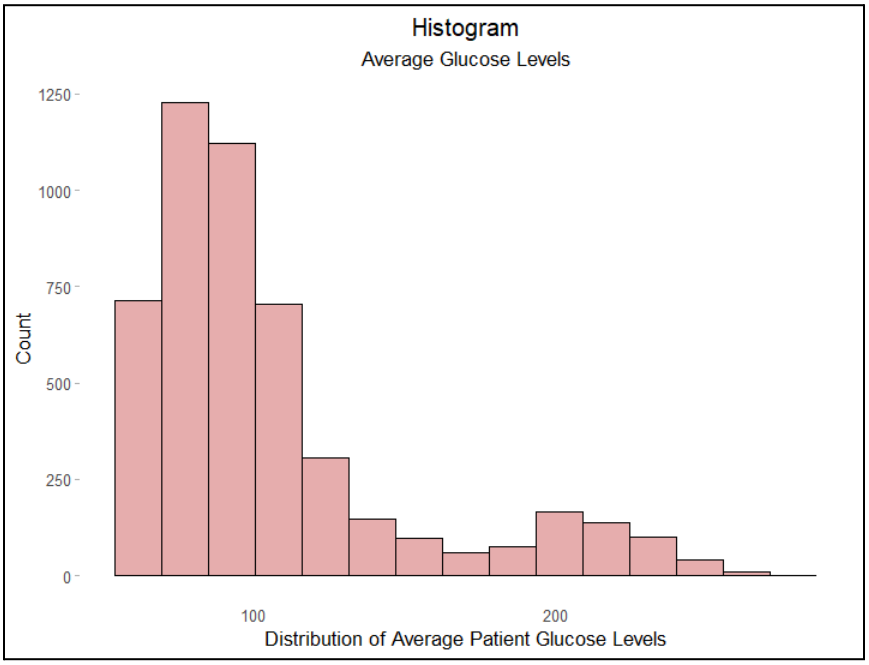


Figure 6: Histogram for the distribution of patients based on average glucose level. The distribution of these values is not normal. The heavy right skew indicates a number of patients with abnormally high glucose levels.

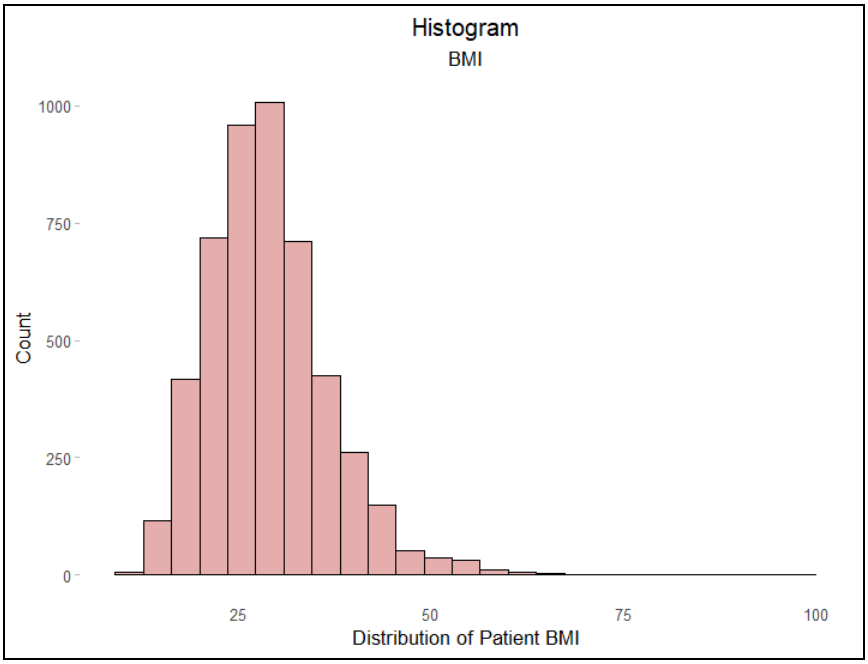
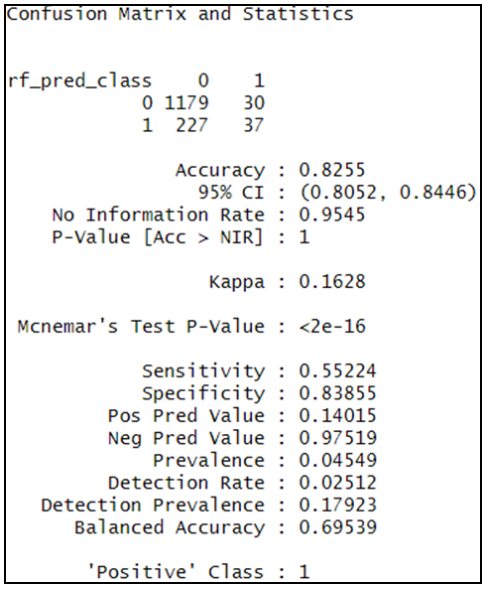


Figure 7: Histogram for the distribution of patients based on BMI.

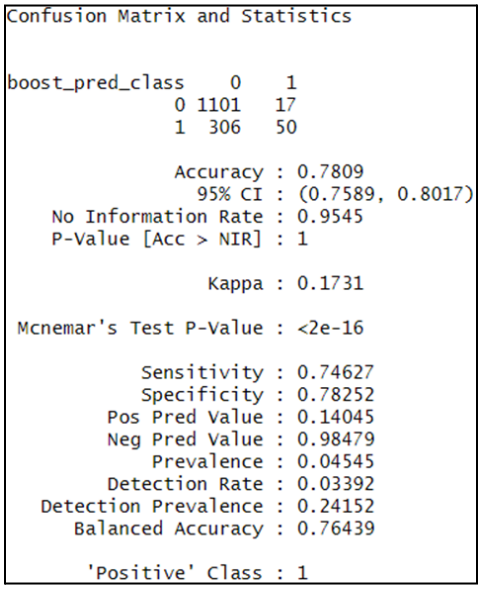
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1232   34
         1  174   33

               Accuracy : 0.8588
                 95% CI : (0.84, 0.8762)
    No Information Rate : 0.9545
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1849

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

            Sensitivity : 0.49254
            Specificity : 0.87624
         Pos Pred Value : 0.15942
         Neg Pred Value : 0.97314
             Prevalence : 0.04549
         Detection Rate : 0.02240
   Detection Prevalence : 0.14053
      Balanced Accuracy : 0.68439

       'Positive' Class : 1
```

Figure 8: Confusion matrix for logistic regression.



```
Confusion Matrix and Statistics


rf_pred_class    0    1
            0 1179   30
            1  227   37

               Accuracy : 0.8255
                 95% CI : (0.8052, 0.8446)
    No Information Rate : 0.9545
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1628

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

            Sensitivity : 0.55224
            Specificity : 0.83855
         Pos Pred Value : 0.14015
         Neg Pred Value : 0.97519
             Prevalence : 0.04549
         Detection Rate : 0.02512
   Detection Prevalence : 0.17923
      Balanced Accuracy : 0.69539

       'Positive' Class : 1
```

Figure 9: Confusion matrix for random forest.



```
Confusion Matrix and Statistics


boost_pred_class    0    1
               0 1101   17
               1  306   50

               Accuracy : 0.7809
                 95% CI : (0.7589, 0.8017)
    No Information Rate : 0.9545
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1731

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

            Sensitivity : 0.74627
            Specificity : 0.78252
         Pos Pred Value : 0.14045
         Neg Pred Value : 0.98479
             Prevalence : 0.04545
         Detection Rate : 0.03392
   Detection Prevalence : 0.24152
      Balanced Accuracy : 0.76439

       'Positive' Class : 1
```

Figure 10: Confusion matrix for XGBoost.

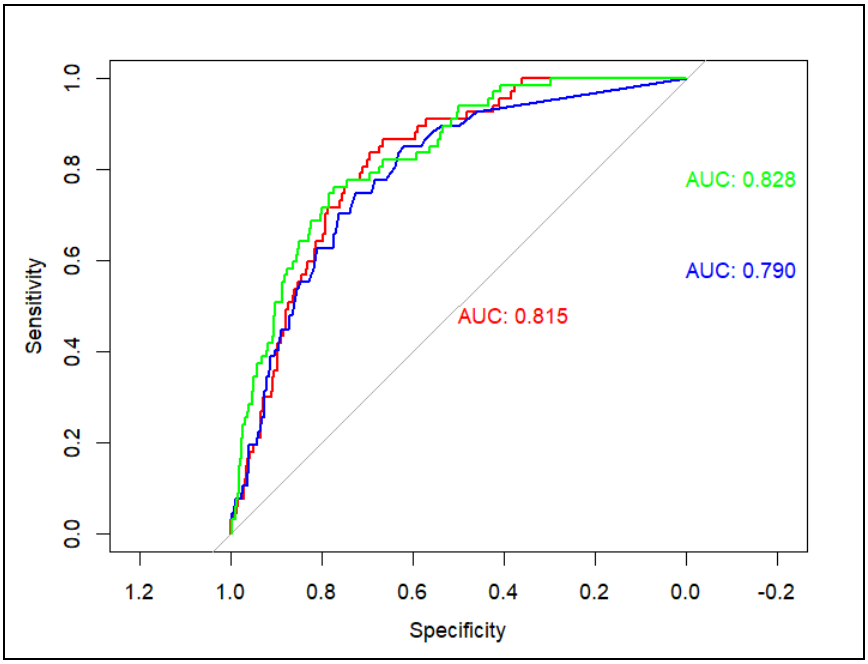


Figure 11: ROC plot for model prediction. Red is logistic regression, blue is random forest, and green is XGBoost.

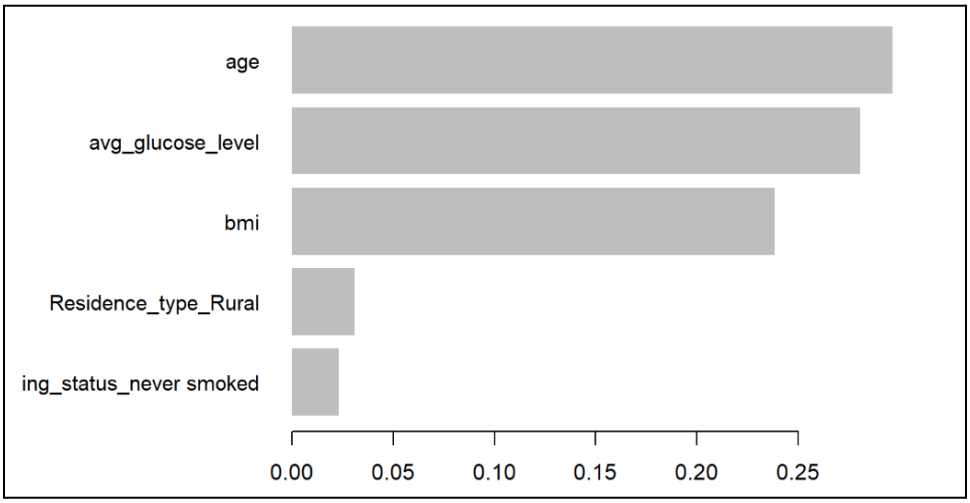


Figure 12: Variable importance chart for the XGBoost model. According to the plot, the top 3 most impactful variables in predicting stroke risk are age, average glucose level, and BMI.