COMP 562 Final Project Report

Andrew Shooman, Ameer Qaqish, Hugh Williamson, Jack Herbert

May 15, 2021

1 Introduction

According to the CDC, stroke is the fifth leading cause of death for Americans. It accounts for 1 in every 6 deaths from cardiovascular disease in the US. Additionally, nearly 795,000 people in the US experience a stroke each year. About 77 percent of these are first time strokes, meaning awareness of risk can help to prevent many of these occurrences. For our project we chose to investigate the characteristics and variables that are associated with susceptibility to stroke. By identifying the characteristics that are associated with higher risk of stroke through different models, we hope to gain insight on how to predict strokes. This data is especially useful for practitioners or even the individuals themselves to assess susceptibility to stroke. Addressing the specific variables that are identified as indicators could prevent a significant number of first time cases we see today. Recommendations based on our results could also save practitioners, patients, and individuals time and money.

2 Relevant Work

Classifying rare events is a difficult problem that is especially prevalent in the medical field. Occurrence of stroke is an example of such a rare event. We examine a dataset to gain insight on what models do well in predicting strokes.

3 Approach

3.1 The Dataset

The dataset we examined has 5110 samples and has the following variables:

Figure 1: Summary of variables in the dataset.

```
   gender          age         hypertension heart_disease ever_married         work_type     Residence_type
 Female:2995   Min.   : 0.08   0:4612       0:4834        No :1757     children     : 687   Rural:2514
 Male  :2115   1st Qu.:25.00   1: 498       1: 276        Yes:3353     Govt_job     : 657   Urban:2596
               Median :45.00                                           Never_worked :  22
               Mean   :43.23                                           Private      :2925
               3rd Qu.:61.00                                           Self-employed: 819
               Max.   :82.00
 avg_glucose_level      bmi                 smoking_status stroke
 Min.   : 55.12    Min.   :10.30    formerly smoked: 885   no :4861
 1st Qu.: 77.25    1st Qu.:23.70    never smoked   :1892   yes: 249
 Median : 91.89    Median :28.30    smokes         : 789
 Mean   :106.15    Mean   :28.94    Unknown        :1544
 3rd Qu.:114.09    3rd Qu.:32.93
 Max.   :271.74    Max.   :97.60
```

We used the stroke variable as the outcome, and the rest of the variables as covariates in our analysis. In the dataset, 4% of the samples had a missing value for bmi. We imputed these values using a linear regression of bmi on the rest of the variables. Our goal was to find a model that can accurately predict the stroke variable from the other variables.

3.2 Model Fitting

We split the data set into a training set containing 80% of the data, and a test set containing 20% of the data. The split was done randomly. We fit the following models on the training set: logistic regression, support vector machine with linear kernel, k-nearest neighbors, random forest, and stochastic gradient boosting.

Since only 5% of the samples had a stroke, we chose to use Cohen's unweighted Kappa statistic as a metric for choosing the tuning parameters for all the models we fitted. We now explain precisely how we chose the hyperparameters for each model. Logistic regression had no hyper-parameters. For the support vector machine, we chose the cost hyperparameter to be $C = 1$. For the k-nearest neighbors, we chose the k hyperparameter from the set $\{5, 7, 9\}$. For each $k \in \{5, 7, 9\}$ we did cross validation with 5 folds repeated 3 times and computed the average value of the Kappa statistic when making predictions using the k-nearest neighbor algorithm with k neighbors using a threshold of 0.5 for acceptance (meaning sample i is classified as having a stroke when the probability p_i outputted by the model for this sample is greater than 0.5). We chose k for which the k-nearest neighbor produced algorithm produced the largest average value of the Kappa statistic. This was $k = 5$. For the random forest model, we used 500 trees. For the number m of variables randomly sampled as candidates at each split, we chose m from the set $\{2, 8, 15\}$ using the same procedure as for choosing k in the k-nearest neighbors algorithm. We chose $m = 15$. For the stochastic gradient boosting model hyperparameters, we chose the max tree depth from $\{1, 2, 3\}$, number of trees from $\{50, 100, 150\}$, a shrinkage value of 0.1, and a minimum terminal node size of 10. We chose the max tree depth to be 2 and the number of trees to be 150 using the same procedure as for choosing k in the k-nearest neighbors algorithm.

After each model was trained, the output of each model was a probability p_i for each sample to have a stroke. In order to make predictions, for each model we chose a threshold probability t and considered $p_i > t$ as a prediction that sample i had a stroke, and $p_i \leq t$ as a prediction that sample i did not have a stroke. For each model (so different models had different thresholds), the threshold t was chosen to maximize Cohen's unweighted Kappa statistic when doing prediction on the training set.

4 Results

The logistic regression model coefficients are displayed in figure 2.

Figure 2: Logistic regression model coefficients.

```
Coefficients:
                (Intercept)                      genderMale                          age
                 -7.1990564                      -0.0299908                    0.0708767
              hypertension1                  heart_disease1               ever_marriedYes
                  0.5312954                       0.1367771                   -0.2434567
         work_typeGovt_job            work_typeNever_worked            work_typePrivate
                 -0.1787615                      -9.7633996                   -0.0055803
    `work_typeSelf-employed`             Residence_typeUrban            avg_glucose_level
                 -0.3638941                       0.0407745                    0.0040606
                        bmi   `smoking_statusnever smoked`         smoking_statussmokes
                  0.0007275                      -0.2865638                    0.1172175
      smoking_statusUnknown
                 -0.0087483
```

From these coefficients, we see that having hypertension, having heart disease, and smoking were the most influential covariates that were positively associated with stroke according to the model. We also see that never working, never smoking, and having ever been married were the most influential covariates that were negatively associated with stroke. We ran each model on the test set. The resulting confusion matrices for each of the models are displayed in figures 3, 4, 5, 6, 7.

5 Conclusion

From the confusion matrices, we see that all the models had relatively poor sensitivity compared to specificity, that is, the models struggled to classify those who had a stroke as having a stroke. This

Figure 3: **Logistic Regression**

```
          Reference
Prediction  no yes
       no  875  24
       yes  97  25

               Accuracy : 0.8815
                 95% CI : (0.8601, 0.9007)
    No Information Rate : 0.952
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2404
```

Figure 4: **Support vector machine**

```
          Reference
Prediction  no yes
       no  705  27
       yes 267  22

               Accuracy : 0.712
                 95% CI : (0.6832, 0.7397)
    No Information Rate : 0.952
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0524
```

is probably due to the low proportion of positive examples in the dataset. Out of all the models, the logistic regression had the highest sensitivity and kappa statistic, although the logistic regression model had the second highest number of false positive predictions after the svm model.

6 Further Considerations

The classification of strokes was difficult mainly due to the low number of positive examples in the dataset. Obtaining more data and using more covariates could help improve the accuracy of the models. Different ways of choosing the hyper-parameters may also give better results.

One thing we observed in the data set is that the 4% of individuals with missing bmi values had strokes at a higher proportion (20%) than those with bmi values. Since the source of the data is confidential, we do not know the reason for this.

References

[1] Stroke Prediction Dataset,
`https://www.kaggle.com/fedesoriano/stroke-prediction-dataset`

Figure 5: ***k*-Nearest Neighbors**

```
          Reference
Prediction  no yes
       no  926  38
       yes  46  11

               Accuracy : 0.9177
                 95% CI : (0.8992, 0.9338)
    No Information Rate : 0.952
    P-Value [Acc > NIR] : 1.000

                  Kappa : 0.1644
```

Figure 6: **Random forest**

```
          Reference
Prediction  no yes
       no  925  41
       yes  47   8

               Accuracy : 0.9138
                 95% CI : (0.8949, 0.9303)
    No Information Rate : 0.952
    P-Value [Acc > NIR] : 1.000

                  Kappa : 0.1086
```

Figure 7: **Stochastic Gradient Boosting**

```
          Reference
Prediction  no yes
       no  902  30
       yes  70  19

               Accuracy : 0.9021
                 95% CI : (0.8822, 0.9196)
    No Information Rate : 0.952
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2275
```