Stroke Prediction under Class Imbalance: A Comparison of Statistical and Machine Learning Models

Jiwoo Ha

1 Introduction

1.1 Data Description

The cerebral stroke dataset used in this study consists of 43,400 observations and includes 11 predictor variables along with one categorical response variable. The dataset was obtained from Kaggle (https://www.kaggle.com/datasets/viviansam/cerebral-stroke-dataset). Among the predictors, eight are categorical variables: *id, gender, hypertension, heart disease, ever married, work type, residence type,* and *smoking status*. The remaining three predictors: *age, average glucose level,* and *BMI*, are numerical variables. The response variable, *stroke*, indicates the occurrence of a stroke and is categorical.

The variable *gender* consists of three categories: Female, Male, and Other. The variables *hypertension*, *heart disease*, and *ever married* are binary, coded as 0 (No) and 1 (Yes). The *work type* variable includes five categories: Children, Private, Never worked, Self-employed, and Government job. *Residence type* has two categories: Rural and Urban. The variable *smoking status* includes three categories: Never smoked, Formerly smoked, and Smokes.

Regarding the numerical predictors, *age* ranges from 0.08 to 82.0 years. The *average glucose level* ranges from 55.0 to 291.05, whereas *BMI* ranges from 10.1 to 97.6.

1.2 Data Preprocessing

The dataset originally included an *id* field, which served only as an identifier and carried no predictive information. This variable was removed prior to analysis, leaving 10 predictor variables for modeling.

1.2.1 Missing Data Handling

For the *gender* variable, 11 observations were labeled as “Other.” Given the large

sample size of 43,400 observations, these cases were considered negligible and were subsequently removed from the dataset.

In the *smoking status* variable, approximately 30% of the data were missing. Upon stratifying the dataset into children (<18 years) and adults (≥18 years), a substantial discrepancy emerged. 81% of children had missing smoking information, while only about 20% of adults had missing values. Since smoking patterns differ intrinsically between these age groups—and children are expected to have essentially near-zero smoking prevalence—two separate imputation categories were created: "Unknown_child" and "Unknown_adult".

For the *BMI* variable, no significant gender differences were observed. However, notable variation existed across age groups. Children (<13 years), teens (<18 years), and adults (≥18 years) had median BMI values of 18.4, 23.5, and 29.1, respectively. Therefore, *BMI* was imputed using the median value within each age group to better preserve the underlying distributional differences.

1.2.2 SMOTE (Synthetic Minority Oversampling Technique)

The response variable in this dataset exhibits a pronounced class imbalance, with only approximately 1.8% of the observations corresponding to stroke cases. To address this imbalance and improve model performance, the Synthetic Minority Oversampling Technique (SMOTE) was employed. Unlike simple random oversampling which merely duplicates minority-class observations, SMOTE generates synthetic samples by interpolating between minority-class instances and their k nearest neighbors, thereby promoting a more representative decision boundary.

As the dataset contains both categorical and continuous predictors, applying the

standard version of SMOTE could result in invalid categorical values, such as *hypertension* = 0.5. To circumvent this issue, I utilized SMOTE-NC, an adaptation of SMOTE specifically designed for mixed-type datasets. SMOTE-NC ensures that categorical variables are handled appropriately by generating only legitimate category labels, while continuous variables undergo conventional interpolation.

The SMOTE-NC procedure was applied exclusively to the training set to prevent data leakage; the dataset had been partitioned into training and test sets using an 80/20 split. After oversampling, the continuous variables were standardized based on the statistics of the original training set, and this scaling was subsequently applied to both the oversampled training set and the test set, prior to model fitting.

2 Models

2.1 Logistic Regression

Logistic regression was selected as a natural baseline model for this analysis, given the binary structure of the stroke outcome variable. The model offers clear interpretability, allowing us to assess the influence of individual risk factors on stroke occurrence through its coefficient estimates. In addition, logistic regression is computationally efficient and can be trained rapidly even on large datasets, making it an appropriate starting point before considering more complex nonlinear models.

2.1.1 Additional Terms for Logistic Regression

To allow the logistic regression model to capture clinically meaningful joint effects, I incorporated several interaction terms among the predictors. These included *hypertension* × *average glucose level*, *age* × *hypertension*, *BMI* × *age*, and *BMI* × *average glucose*

level, which were motivated by the possibility that the combined influence of these risk factors may exceed their individual contributions. I also added an *age* × *is_smoker* interaction to reflect the greater cumulative exposure experienced by older smokers. The smoking-status indicator, *is_smoker*, originally introduced only to construct this interaction, remained in the final model because it demonstrated meaningful predictive value on its own.

In addition to the interaction terms, I included a squared BMI term (BMI^2) to allow the model to represent potential nonlinear associations between BMI and stroke risk. Given that the relationship between adiposity and cerebrovascular outcomes is often not strictly linear—particularly at very low or very high BMI levels—incorporating a polynomial term provided a simple way to capture such curvature in the response surface.

Additional terms were created only for the logistic regression dataset after applying SMOTE-NC and before scaling, so that the standardization reflected all predictors used in that model.

2.2 Tree-based Machine Learning Models

Tree-based ensemble methods were included in our modeling framework because they naturally accommodate nonlinear relationships and interactions without requiring explicit feature engineering. They also work well with mixed data types and tend to perform strongly on tabular clinical datasets. Below, I briefly summarize the intuition behind each method and the reasons for including them in our analysis.

2.2.1 Random Forest

Random Forest is an ensemble method that constructs many decision trees using

bootstrap samples of the data and aggregates their predictions through majority voting. Because each tree captures different aspects of the training data, the combined model becomes more stable and less prone to overfitting than a single decision tree. Random Forest also handles both numerical and categorical variables with minimal preprocessing and provides feature importance measures, which allow us to interpret the relative influence of individual predictors on stroke risk.

2.2.2 Gradient Boosting

Gradient boosting models build decision trees in a sequential manner, where each new tree is trained to correct the errors made by the previous ones. This iterative refinement process often leads to highly accurate models, and gradient boosting has become one of the most successful techniques for structured/tabular data. I include three commonly used implementations—XGBoost, LightGBM, and CatBoost—to compare their performance and behavior on our dataset.

XGBoost serves as a strong benchmark due to its robust regularization framework and consistent predictive performance across a wide range of classification problems.

LightGBM was selected for its computational efficiency. Its leaf-wise tree growth strategy allows the model to train significantly faster than traditional gradient boosting approaches, and it offers built-in mechanisms for handling categorical variables, which reduces preprocessing effort.

CatBoost is specifically designed for datasets with many categorical features. Its "ordered boosting" procedure helps prevent target leakage and typically improves generalization, making it a suitable option for our dataset, which includes several categorical predictors.

3 Result

3.1 Feature Importance

3.1.1 Logistic Regression

```
> summary(step_model_bic)

Call:
glm(formula = stroke ~ heart_disease + age + bmi_glucose, family = binomial,
    data = train)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.20629    0.08958 -58.118  < 2e-16 ***
heart_disease1  0.72943    0.10294   7.086 1.38e-12 ***
age             1.62190    0.07149  22.686  < 2e-16 ***
bmi_glucose     0.14740    0.03257   4.526 6.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6276.4  on 34720  degrees of freedom
Residual deviance: 5187.5  on 34717  degrees of freedom
AIC: 5195.5

Number of Fisher Scoring iterations: 8
```

Figure 1: Logistic regression model output selected by BIC in R

Since the SMOTE-applied dataset contains synthetic samples, I first evaluated variable significance using the original (non-SMOTE) training data. When I performed stepwise selection using BIC, only *age*, *heart disease*, and the *BMI* × *average glucose level* interaction term were retained, as shown in Figure 1. However, *BMI* × *average glucose level* was not statistically significant in the full model and showed no clear evidence of contributing meaningfully to stroke prediction. In addition, under the principle of marginality, an interaction term should not be included without its corresponding main effects. Yet the main effect of *average glucose level* was also non-significant in the full model. For these reasons, the *BMI* × *average glucose level* interaction was excluded from the final logistic regression model.

```
> summary(full_model)

Call:
glm(formula = formula_logit, family = binomial, data = train)

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -4.82244    0.74099  -6.508 7.61e-11 ***
genderMale                    0.04901    0.08589   0.571  0.56822
genderOther                 -11.22414  765.05661  -0.015  0.98829
hypertension1                 1.17231    0.64063   1.830  0.06726 .
heart_disease1                0.65763    0.10469   6.282 3.35e-10 ***
ever_marriedYes              -0.12070    0.13816  -0.874  0.38231
work_typeGovt_job            -0.82197    0.79801  -1.030  0.30300
work_typeNever_worked        -10.31163  188.03965  -0.055  0.95627
work_typePrivate             -0.63356    0.79157  -0.800  0.42349
work_typeSelf-employed       -0.70731    0.79768  -0.887  0.37524
Residence_typeUrban           0.02946    0.08266   0.356  0.72153
smoking_statusformerly smoked -0.13975    0.12295  -1.137  0.25568
smoking_statusnever smoked    -0.14626    0.11618  -1.259  0.20805
smoking_statussmokes          0.84404    0.53008   1.592  0.11132
is_smoker1                         NA         NA      NA       NA
age                           2.47659    0.37099   6.676 2.46e-11 ***
bmi                           1.62208    0.54042   3.002  0.00269 **
avg_glucose_level             0.08485    0.17179   0.494  0.62138
bmi_square                   -1.14278    0.41527  -2.752  0.00592 **
ht_glucose                   -0.03413    0.06467  -0.528  0.59770
age_ht                       -0.20777    0.15913  -1.306  0.19166
bmi_age                      -0.92951    0.41010  -2.267  0.02342 *
bmi_glucose                   0.11832    0.22271   0.531  0.59524
age_smokes                   -0.21017    0.14392  -1.460  0.14420
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6276.4  on 34720  degrees of freedom
Residual deviance: 5146.4  on 34698  degrees of freedom
AIC: 5192.4

Number of Fisher Scoring iterations: 15
```

Figure 2: Full logistic regression model output using the original data in R

The summary of the full model in Figure 2 also indicates that the *BMI*, BMI^2, and *BMI* × *age* terms have p-values below 0.05, providing justification for their inclusion. Notably, when only *BMI* was included in the model, both *BMI* and *BMI* × *age* were non-significant; however, after adding the BMI^2 term, both effects became significant, indicating that the nonlinear specification better captures the relationship between *BMI* and stroke risk.

Hypertension had a p-value of 0.06726, which is near the conventional threshold and therefore was retained. Additionally, within the *smoking status* variable, the "smokes" category had a p-value of 0.11132, which suggested that collapsing categories could potentially strengthen this effect. Therefore, I considered *is_smoker* reasonable to include as well.

Hence, the final model specified as *stroke* ~ *age* + *hypertension* + *heart disease* + *is_smoker* + *BMI* + BMI^2 + *BMI* × *age*.

```
> summary(final_model1)

Call:
glm(formula = stroke ~ age + hypertension + heart_disease + is_smoker +
    bmi + bmi_square + bmi_age, family = binomial, data = train)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.38423    0.11507 -46.789  < 2e-16 ***
age             2.25940    0.35279   6.404 1.51e-10 ***
hypertension1   0.29165    0.09786   2.980  0.00288 **
heart_disease1  0.72503    0.10320   7.026 2.13e-12 ***
is_smoker1      0.21593    0.11093   1.947  0.05159 .
bmi             1.54932    0.52498   2.951  0.00317 **
bmi_square     -1.06146    0.38794  -2.736  0.00622 **
bmi_age        -0.75056    0.39470  -1.902  0.05722 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6276.4  on 34720  degrees of freedom
Residual deviance: 5183.1  on 34713  degrees of freedom
AIC: 5199.1

Number of Fisher Scoring iterations: 9
```

Figure 3: Final logistic regression model output using the original training data in R

The final logistic regression model fitted on the original training dataset showed that all selected predictors were either significant or borderline significant, as shown in Figure 3. This suggests that *age*, *hypertension*, *heart disease*, smoking indicator *is_smoker*, and both the linear and nonlinear *BMI* terms all contribute meaningfully to stroke risk in the original dataset.

```
> summary(final_model2)

Call:
glm(formula = stroke ~ age + hypertension + heart_disease + is_smoker +
    bmi + bmi_square + bmi_age, family = binomial, data = train_balanced)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.95257    0.02596 -75.212   <2e-16 ***
age             2.74243    0.09258  29.621   <2e-16 ***
hypertension1   0.29875    0.02654  11.258   <2e-16 ***
heart_disease1  0.61367    0.03249  18.889   <2e-16 ***
is_smoker1      0.36108    0.02656  13.594   <2e-16 ***
bmi             4.26382    0.14103  30.234   <2e-16 ***
bmi_square     -3.75852    0.11620 -32.345   <2e-16 ***
bmi_age        -0.90946    0.10405  -8.741   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 94529  on 68187  degrees of freedom
Residual deviance: 60590  on 68180  degrees of freedom
AIC: 60606

Number of Fisher Scoring iterations: 6
```

Figure 4: Final logistic regression model output using the oversampled training data in R

Because the training data contained very few true stroke cases, I trained the logistic regression model using the SMOTE-NC-oversampled dataset to ensure that the minority class was adequately represented. The resulting model, which was later applied to the untouched test dataset, is given by:

$$\text{logit}(\hat{p}) = -1.95257 + 2.74243 \cdot age + 0.29875 \cdot hypertension + 0.61367 \cdot heart\ disease + 0.36108 \cdot is\_smoker + 4.26382 \cdot BMI - 3.75852 \cdot BMI^2 - 0.90946 \cdot BMI \times age.$$

This final model was then used to generate predicted stroke probabilities for the test dataset.

3.1.2 Tree-based Machine Learning Models

```
> print(rf_importance_df[1:10, ])
                            Feature Importance
age                             age 13684.6038
bmi                             bmi  3989.6026
avg_glucose_level avg_glucose_level  3758.5394
ever_married           ever_married  1866.5773
work_type                 work_type  1659.1228
smoking_status       smoking_status  1471.7531
heart_disease         heart_disease   782.8572
gender                       gender   706.4376
Residence_type       Residence_type   694.2543
hypertension           hypertension   613.9765
```

Figure 5: Feature Importance in Random Forest

```
> print(xgb_importance[1:10, ])
              Feature        Gain       Cover   Frequency
               <char>       <num>       <num>       <num>
 1:               age 0.852175378 0.683137248 0.33641161
 2:               bmi 0.049280425 0.112603898 0.19591029
 3: avg_glucose_level 0.031171830 0.091426670 0.19261214
 4:      ever_married 0.016493162 0.014967138 0.01649077
 5:         work_type 0.015973716 0.045019544 0.07387863
 6:    smoking_status 0.011859047 0.011828200 0.08509235
 7:     heart_disease 0.009550501 0.026886387 0.01583113
 8:      hypertension 0.005370081 0.007407263 0.02242744
 9:            gender 0.005240782 0.004074966 0.03562005
10:    Residence_type 0.002885079 0.002648688 0.02572559
```

Figure 6: Feature Importance in XGBoost

```
> print(lgb_importance[1:10, ])
              Feature        Gain       Cover   Frequency
               <char>       <num>       <num>       <num>
 1:               age 0.798656051 0.771950104 0.758333333
 2:      ever_married 0.047838377 0.017999186 0.004166667
 3:         work_type 0.039424854 0.036168779 0.031250000
 4:               bmi 0.029049072 0.057593901 0.058333333
 5:     heart_disease 0.025932338 0.022500960 0.006250000
 6: avg_glucose_level 0.024691376 0.049211168 0.056250000
 7:      hypertension 0.018291785 0.021555647 0.025000000
 8:    smoking_status 0.010810187 0.015899259 0.035416667
 9:            gender 0.003622669 0.004909169 0.016666667
10:    Residence_type 0.001683290 0.002211829 0.008333333
```

Figure 7: Feature Importance in LightGBM

```
> print(cat_importance_df[1:10, ])
                            Feature Importance
age                             age  72.268757
bmi                             bmi  20.207782
avg_glucose_level avg_glucose_level   5.360280
hypertension           hypertension   1.092344
heart_disease         heart_disease   1.070837
gender                       gender   0.000000
ever_married           ever_married   0.000000
work_type                 work_type   0.000000
Residence_type       Residence_type   0.000000
smoking_status       smoking_status   0.000000
```

Figure 8: Feature Importance in CatBoost

Across all tree-based machine-learning models, *age* consistently emerged as the most

influential predictor. Its dominance across both linear and nonlinear models suggests a strong, largely monotonic relationship with stroke risk—one that does not require complex modeling to be detected. *BMI* also appeared repeatedly among the top predictors. Whereas logistic regression required additional nonlinear terms (BMI^2 and $BMI \times age$) to capture its effect, tree-based models identified its importance without such transformations, indicating that *BMI* contributes meaningfully in ways that extend beyond simple linear effects. *Average glucose level* showed moderate importance in most tree-based models but was not strongly significant in logistic regression, suggesting that its association with *stroke* risk may involve threshold effects or nonlinear patterns that linear models are not well equipped to capture.

In addition to these continuous predictors, several categorical variables, most notably *ever_married* and *work_type*, ranked within the top five features in models such as XGBoost and LightGBM, despite being non-significant in logistic regression. Their elevated importance in nonlinear models implies that these variables may influence stroke risk through interactions or complex structures that logistic regression does not detect.

Taken together, the top predictors across machine-learning models indicate that stroke risk is driven primarily by *age*, *BMI*, and *average glucose level*, along with several sociodemographic factors whose effects appear nonlinear and are therefore better captured by flexible, ensemble-based methods than by logistic regression.

3.2 Confusion Matrix

```
> print(lr_cm)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6341   30
         1 2182  126

               Accuracy : 0.7451
```

```
> print(rf_cm)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7679   97
         1  844   59

               Accuracy : 0.8916
```

Figure 9: Confusion Matrix for logistic regression (left) and random forest (right)

```
> print(xgb_cm)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7010   57
         1 1513   99

               Accuracy : 0.8191
```

```
> print(lgb_cm)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7616   83
         1  907   73

               Accuracy : 0.8859
```

Figure 10: Confusion Matrix for XGBoost (left) and LightGBM (right)

```
> print(cat_cm)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6183   26
         1 2340  130

               Accuracy : 0.7274
```

Figure 11: Confusion Matrix of CatBoost

Across all models, accuracy alone was misleading due to the extreme class imbalance, so the more meaningful comparison lies in how well each model recovered actual stroke cases—that is, their sensitivity. Logistic regression showed high sensitivity, identifying 126 stroke cases, but at the cost of a large number of false positives. Random Forest achieves the highest overall accuracy but has the lowest sensitivity among all models, capturing only 59 stroke cases and therefore failing to meaningfully detect the true strokes. XGBoost performs better than Random Forest, identifying 99 stroke cases, but it is still hard to say its performance is

considered good. LightGBM also improves upon Random Forest, retrieving 73 stroke cases, though it remains less sensitive than XGBoost. CatBoost identifies the largest number of stroke cases, which is 130, showing the highest sensitivity; however, this comes with a large increase in false positives. Overall, these results highlight that while most models classify non-stroke cases well, their ability to recover true stroke cases varies considerably, with logistic regression and CatBoost performing best in terms of sensitivity.

3.3 Evaluation Metrics

```
> print(results)
                Model    AUC F1_Score Sensitivity Specificity
1 Logistic Regression 0.8489   0.1023      0.8077      0.7440
2       Random Forest 0.8068   0.0983      0.2949      0.9139
3             XGBoost 0.8355   0.1080      0.6410      0.8129
4            LightGBM 0.8421   0.1418      0.4359      0.9138
5            CatBoost 0.8415   0.1012      0.8205      0.7366
```

Figure 12: Evaluation Metrics for all models

The evaluation metrics illustrate how differently the models perform depending on the aspect of classification being emphasized. Logistic regression achieves the highest AUC (0.8489), indicating strong overall separation between classes despite the imbalance. CatBoost attains the highest sensitivity (0.8205), meaning it identifies the largest proportion of true stroke cases, though its specificity is comparatively low. Random Forest, while showing high specificity (0.9139), has the lowest sensitivity (0.2949), confirming that it predominantly predicts the majority class and misses most stroke cases. LightGBM has the highest F1 score but it is all because it detected true negative well, not because it detected true positive. As for XGBoost, its sensitivity is moderate and its computational time is relatively long, making it difficult to justify its use given that it does not clearly outperform the other models.

4 Conclusion

Because early detection of stroke is far more important than avoiding false positives, sensitivity

is the primary metric of interest in this context, rather than the F1-score or specificity. From a computational and cost perspective, logistic regression is the most efficient model, and given that it also achieves the highest AUC and relatively high sensitivity, it is the most practical choice when resources are limited. However, if computational cost is not a concern, CatBoost would be the preferred model because of its superior sensitivity.

However, during the logistic regression modeling process, I did not incorporate the nonlinear pattern in *average glucose level* that the machine learning models were able to capture, which likely prevented the final model from fully leveraging an important predictive signal. Incorporating a transformed term for *average glucose level* may have allowed logistic regression to perform more competitively, and possibly become the best model. In addition, our analysis relied solely on SMOTE for handling class imbalance and did not explore alternative oversampling approaches, such as bootstrap-based methods (e.g., ROSE), which could have provided a more nuanced comparison across models. Future work incorporating these adjustments may provide a more comprehensive assessment of model behavior in this clinical prediction setting.

This project was originally conducted as part of STA 5168: Statistics in Applications III (Fall 2025), and has been expanded for portfolio purposes.