Can demographic and physiological features predict the likelihood of a stroke?

Monirul Islam

Introduction

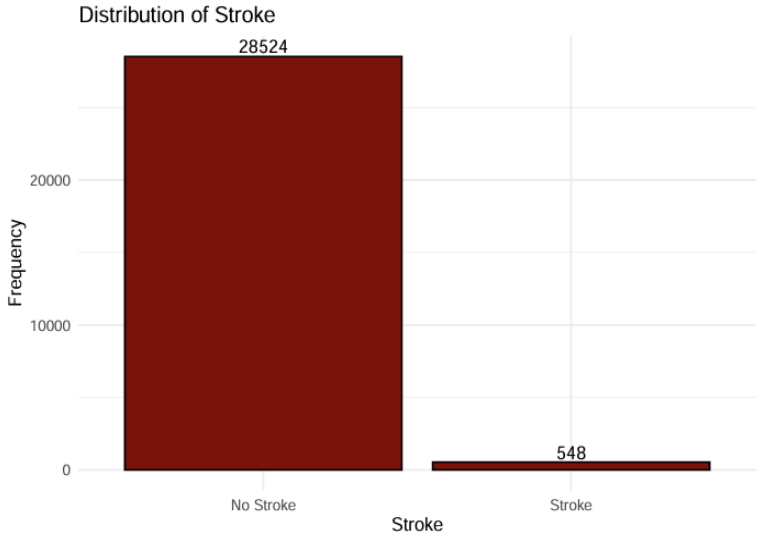
To begin the workflow, the data was first loaded and inspected. It consists of information from 43,400 patients comprising 12 variables with mixed data types. The covariates include gender, age, hypertension status, heart disease status, marital status, work type, residence type, average glucose level, body mass index, and smoking status. The target variable is stroke occurrence. All the covariates mentioned were initially included in a baseline logistic regression model and subsequently refined to achieve optimal predictive performance for the likelihood of a stroke.

Missing Values

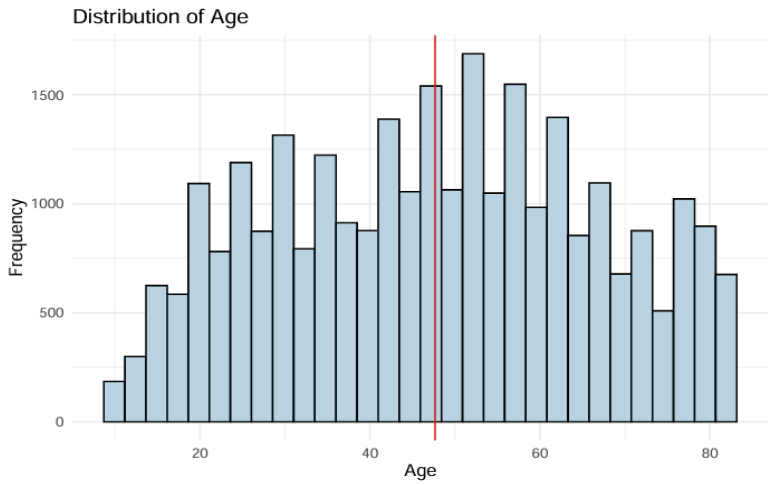
The dataset was screened for missing values after loading. There were 1462 missing values for the body mass index variable and 13,292 missing or empty values for smoking status. After excluding records with missing values, the dataset was reduced to 29,072 observations. The smoking status variable originally comprised three categories: never smoked, formerly smoked, and smokes. This covariate was transformed into a binary variable by combining individuals who currently smoke and or formerly smoked into a new category labeled “ever smoked,” while those who never smoked remained in the existing category.

Exploratory Data Analysis

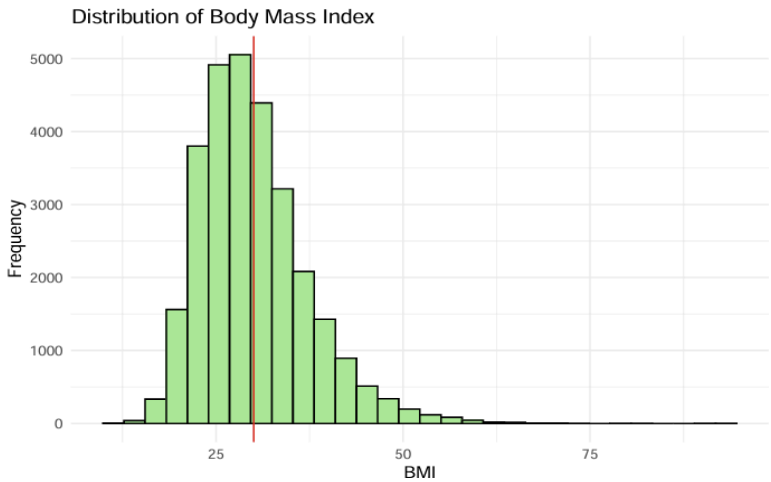
The following visualizations were created to analyze the distributions of selected variables.



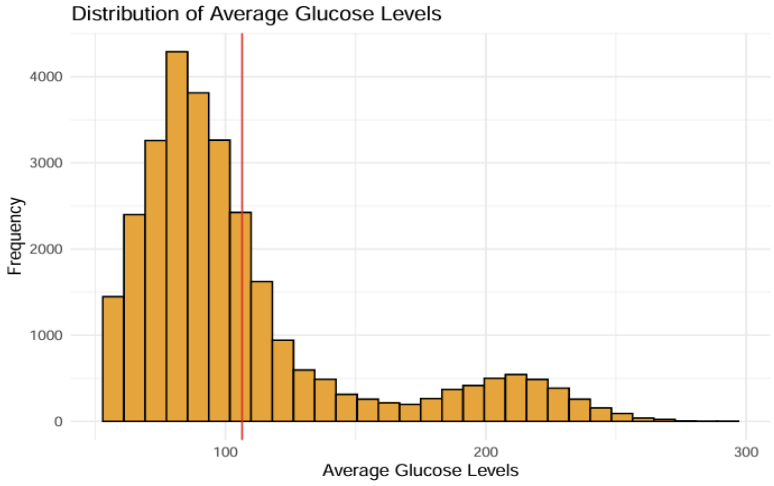
There is a significant uneven distribution of the target variable, with 28,524 cases of no stroke and 548 cases of stroke.



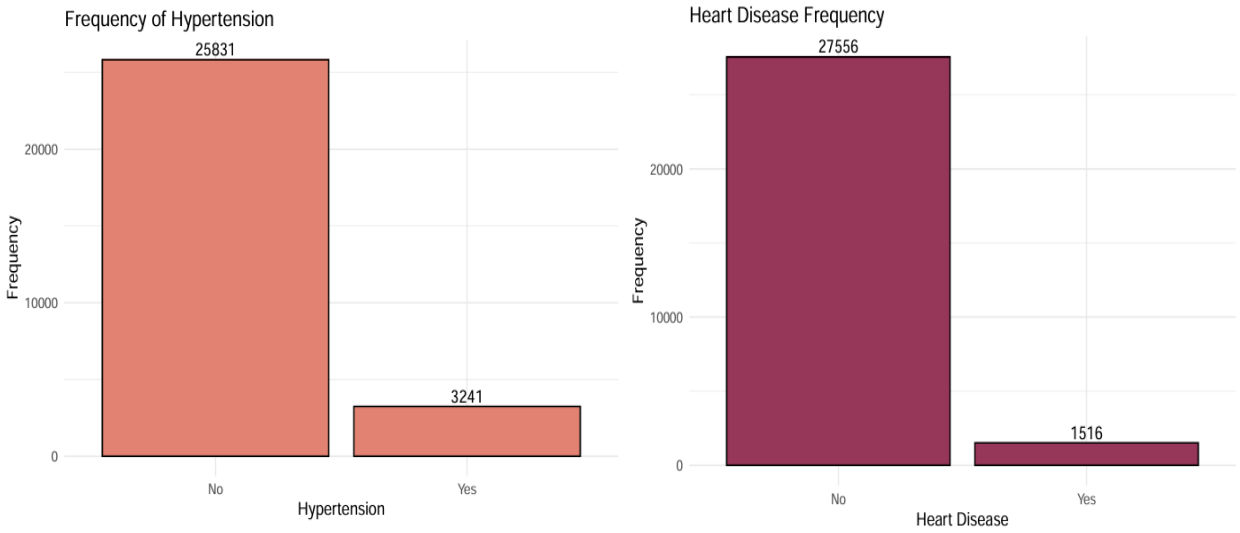
The distribution of age appears fairly symmetric with no extreme skewness. The average age of the patient pool is around the late 40s.



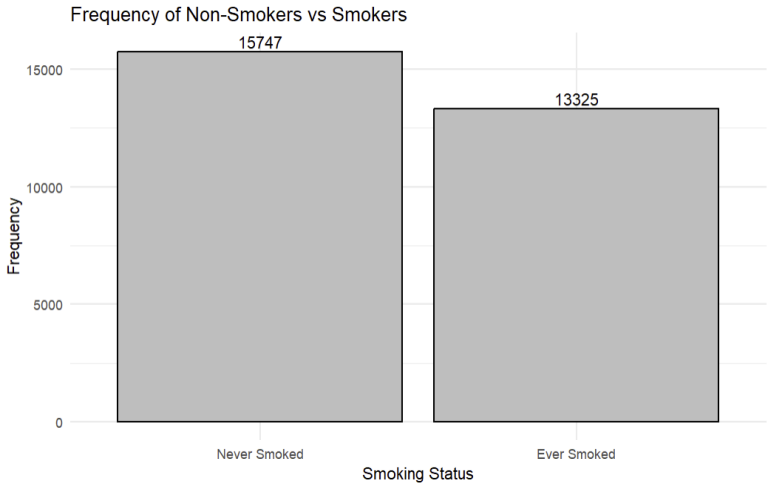
Body mass index is right-skewed, possibly implying the presence of outliers. Extreme values are observable above the 50 mark. The mean lies to the right of the peak.



The distribution of average glucose levels also appears rightly skewed, with potential outliers. There are two concentrated groups, one group with normal glucose levels and another with elevated levels.



There is a heavy imbalance in both the frequency of hypertension and heart disease. The majority of patients in the datasets identify having none of these conditions. Despite low prevalence, both covariates are worth observing to see their influence on stroke risk as standalone factors and as interaction terms in the subsequent model.



After transforming the original smoking status variable, the resulting distribution was more balanced between patients who never smoked and those who currently or formerly smoked.

Base Model

Anova Analysis

```
Analysis of Deviance Table (Type II tests)

Response: stroke
                  LR Chisq Df Pr(>Chisq)
gender                0.07  2     0.9666
age                 439.10  1  < 2.2e-16 ***
hypertension         18.51  1  1.692e-05 ***
heart_disease        27.37  1  1.683e-07 ***
ever_married          1.10  1     0.2940
work_type             1.89  4     0.7560
Residence_type        0.08  1     0.7822
avg_glucose_level    23.82  1  1.056e-06 ***
bmi                   1.67  1     0.1961
smoking_status_new    0.67  1     0.4146
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A base logistic regression model was fitted using all the covariates, followed by a Type II ANOVA test to assess each predictor's significance. Variables including age, hypertension status, heart disease status, and average glucose levels are statistically significant, as their p-values are less than alpha 0.05. Out of all the covariates, age exhibits the largest deviance reduction, making it one of the more dominant predictors of stroke risk. Other predictors such as gender, marital status, work type, residence type, body mass index, and smoking status do not appear to be statistically significant as their p-values are greater than 0.05. Nonetheless, all variables were retained for the baseline model to evaluate model performance and enable comparison with the next model.

Train / Test Split

The dataset was split into training and test sets using a 70/30 proportion. Rather than simple random sampling, stratified sampling was performed using the createDataPartition() function to preserve the original stroke imbalance in both sets.

K-fold Cross Validation

The performance of the base logistic regression model was first evaluated with 10-fold stratified cross-validation on the training set only. All the data pre-processing steps were performed within each fold to prevent any data leakage. Categorical variables (gender, marital status, work type, residence type, hypertension, heart disease, and smoking status) were internally transformed into dummy variables by the logistic model. Continuous predictors (age, body mass index, and average glucose level) were standardized with the means and standard deviations. Due to a severe imbalance of the target variable, a reduced classification threshold of 0.1 was incorporated instead of the standard 0.5. Under the 0.1 cutoff, the base model achieved a high accuracy of 95-96 percent across the folds. Sensitivity values ranged from 3 percent to 29

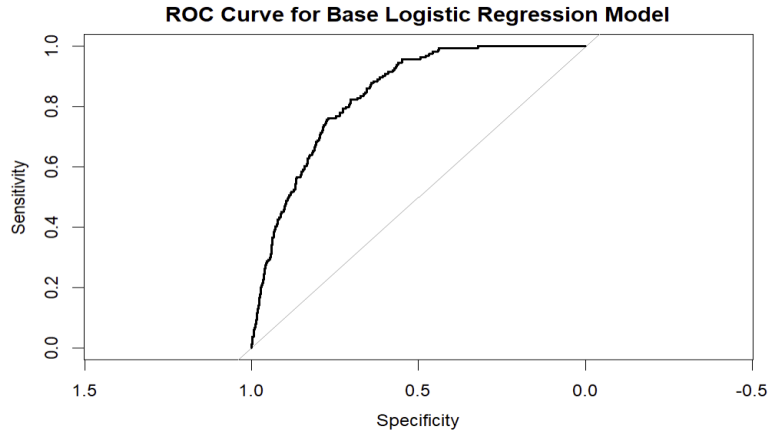
percent. Specificity remained high across the folds, indicating that the majority of non-stroke classes were classified correctly.

Odds Ratio Analysis (Full Training Set)

```
##                                  odds_ratio      p_value
## (Intercept)                    1.060551e-07 9.596195e-01
## genderMale                     9.302878e-01 5.116268e-01
## genderOther                    2.585282e-06 9.959415e-01
## age                            3.822430e+00 2.103234e-58
## hypertensionYes                1.534787e+00 3.990614e-04
## heart_diseaseYes               1.775509e+00 2.949491e-05
## ever_marriedYes                8.491592e-01 3.433940e-01
## work_typeGovt_job              7.967692e+04 9.716163e-01
## work_typeNever_worked          5.431758e-01 9.994018e-01
## work_typePrivate               7.714529e+04 9.716975e-01
## work_typeSelf-employed         7.261436e+04 9.718497e-01
## Residence_typeUrban            1.004322e+00 9.673564e-01
## avg_glucose_level              1.191118e+00 3.434704e-05
## bmi                            9.133008e-01 1.411168e-01
## smoking_status_newEver Smoked 1.041132e+00 7.073156e-01
```

After training on the full training set, the odds ratios were derived by exponentiating the coefficients. The odds ratio of 1.53 for hypertension means that individuals with such a condition have approximately 53 percent higher odds of stroke compared to those without hypertension. The odds ratio for heart disease is 1.78, indicating people with heart disease have approximately 78 percent higher odds of stroke compared to those without the condition, holding other predictors constant. Age is the strongest predictor with an extremely low p-value. The odds ratio suggests that a one standard deviation increase in age is associated with almost 3.8 times higher odds of a stroke. Lastly, glucose levels are also statistically significant, as a one standard deviation increase is associated with 19 percent higher odds of having a stroke.

Test Set Evaluation



The same classification threshold of 0.1 was utilized when evaluating the trained base model on the test set. The standard metrics of accuracy, sensitivity, specificity, and AUC were computed. The model achieved a high accuracy of approximately 96 percent. The model's ability to correctly classify stroke cases (sensitivity) was 16.5 percent, and non-stroke cases (specificity) was 97.5 percent. The ROC analysis further supports these metrics, as the model achieved an AUC of 0.84, indicating good ability to differentiate between stroke and non-stroke cases.

New Model

Anova Analysis

```
Analysis of Deviance Table (Type II tests)

Response: stroke
                          LR Chisq Df Pr(>Chisq)
age                         489.42  1  < 2.2e-16 ***
hypertension                 18.29  1  1.900e-05 ***
heart_disease                29.44  1  5.767e-08 ***
avg_glucose_level            21.86  1  2.939e-06 ***
bmi                           3.04  1   0.081106 .
smoking_status_new            0.40  1   0.529498
hypertension:heart_disease    3.76  1   0.052628 .
age:hypertension              5.68  1   0.017124 *
age:heart_disease             6.28  1   0.012233 *
age:smoking_status_new        8.07  1   0.004504 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A new logistic regression model was explored, retaining the relevant covariates and inserting interaction terms of interest. A Type II ANOVA test was again applied to assess the significance of the predictors. The refined model shows robust effects from the same predictors observed in the base model. Age remains dominant, with hypertension, heart disease, and average glucose level contributing significantly to stroke risk. Interactions between age and hypertension, age and heart disease, and age and smoking status are statistically significant with p-values less than alpha 0.05. Although smoking status is not significant as a standalone predictor, its interaction with age increases the risk of stroke. Overall, these results suggest that age is the primary driver of stroke risk, with cardiovascular risk factors exerting a secondary influence.

Train / Test Split

The dataset was split in a 70/30 ratio to maintain consistency with the process used for the base model.

Odds Ratio Analysis

```
##                                   odds_ratio       p_value
## (Intercept)                      0.004826365 4.429578e-223
## age                              5.141974748  4.066478e-37
## hypertensionYes                  2.936553938  1.766065e-05
## heart_diseaseYes                 3.346941528  9.350319e-04
## avg_glucose_level                1.180858597  7.468145e-05
## bmi                              0.890664807  6.351793e-02
## smoking_status_newEver Smoked    1.551057051  3.303429e-02
## hypertensionYes:heart_diseaseYes 0.623641673  1.019801e-01
## age:hypertensionYes              0.631765259  1.927671e-02
## age:heart_diseaseYes             0.696320766  1.581662e-01
## age:smoking_status_newEver Smoked 0.677439230  1.477408e-02
```

After fitting the new logistic regression model on the full training set, odds ratios were calculated by exponentiating the coefficients. Age shows a more prominent effect than in the base model, with the odds of a stroke increasing almost fivefold. Hypertension and heart disease also exhibit larger odds ratios compared to the base model. The interaction terms between age and hypertension, and age and smoking status, are statistically significant, possibly signaling that these risk factors are age dependent. The odds ratios being less than one for the interaction terms suggest that hypertension and smoking status have stronger effects at younger ages. The overall likelihood of stroke increases with age.

Test Set Evaluation & Model Comparison

```
                     Model  Accuracy Sensitivity Specificity       AUC
1 Base Logistic Regression 0.9598670   0.1646341   0.9751081 0.8396520
2  New Logistic Regression 0.9594083   0.1585366   0.9747575 0.8414306
```

The new model was applied to the test set, and the same metrics of accuracy, sensitivity, specificity, and AUC were computed, enabling comparison with the base model metric results. The accuracy of both the base model and the new model is nearly the same at approximately 96 percent. There is a slight decrease in sensitivity in the new model. Both models performed well in identifying non-stroke cases with a specificity of 97.5 percent. The new model attained a slightly higher AUC, which entails better differentiation.

```
          df      BIC      AIC
base_glm 15 3353.382 3234.569
new_glm  11 3297.428 3210.298
```

In addition to the standard logistic regression metrics, the models were also compared using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), both which are commonly used to judge model fit. A likelihood ratio test was not appropriate here since the models are not nested. Ideally, lower values for AIC and BIC indicate a better model as these criteria balance goodness of fit and complexity. Since the new regression model has a lower AIC, BIC, and a slight improvement in AUC, it is preferred over the base model. It is critical to

note that model performance is sensitive to the classification threshold, among other factors, and that further experimentation is required to improve metrics like sensitivity.

Bootstrap

A bootstrap with 5,000 replications was performed to further analyze the variability of the sensitivity for the new logistic regression model. In each replication, rows of the test set were sampled with replacement, and the sensitivity was calculated per replication. The bootstrap sensitivity mean was 0.158, with a 95 percent confidence interval between 0.103 and 0.22. The low sensitivity rates across the folds and the wide confidence interval range may indicate that the sensitivity could be dependent on the classification threshold and the class imbalance that exists.

Discussion

Although the model achieved high accuracy and specificity scores, the low sensitivity is not sufficient for clinical purposes, as it is imperative to correctly identify positive cases. Several approaches can be taken to further improve the study. The classification threshold of 0.10 was used for the current experimentation; however, alternative cutoffs can be trialed to identify one that best balances sensitivity and specificity or simply maximizes sensitivity. Given the highly unbalanced stroke class, resampling techniques, such as down-sampling the majority class, can be incorporated into the training data. Lastly, more advanced machine learning models can be used to better capture complex patterns and improve results.