Predictive Analytics in Stroke Risk

Mingze Xu, mx2269

1. Objective

According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Considering stroke, the objective of this project is to apply machine learning techniques to develop robust predictive models for stroke occurrence, based on various health and demographic factors. This project underscores the significance of predictive analytics in healthcare, and aims to bridge the gap between data-driven insights and applications in real-world healthcare situations.

2. Material and Methods

2.1. Data Source and Structures

The source of the data is from Kaggle, and the link to the dataset is Stroke Prediction Dataset. The original data contains 5110 observations with 12 attributes related to health and demographic information, and each observation is a human subject. The attribute information is really straightforwardly shown by its name. Data cleaning, wrangling, and normalization are performed, and the process of these are described in Table 1 attached in Appendix. Also, observations containing N/A are removed, as there are only a few of them.

The dataset is splitted into training, validation, and testing subsets using a 5:1:1 ratio, and the training dataset contains 3434 observations, and the validation and testing datasets contain 737 observations each. The training dataset is used for training the models and it is the primary dataset on which the algorithms learn the patterns; the validation dataset is used to provide an unbiased evaluation of the model fitting during the training phase, and to fine-tune model parameters and select the best-performing ones; the testing dataset is used to provide an unbiased evaluation of the final model fits.

Another data manipulation technique is SMOTETomek, addressing the issue of data imbalance, and will be discussed in Section 3.2 (Sensitivity Analysis) of the Discussion section.

2.2. Metrics

Basic metrics accuracy and ROC-AUC scores are certainly measured. In this project, three other metrics are also used. Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives correctly identified by the model, reflecting its ability to detect the

relevant cases. Precision measures the accuracy in predicting positive cases. F1-score is actually a combination of precision and recall, providing a single score balancing both the model's accuracy and completeness.

2.3. Machine Learning Algorithms used for Predictive Analytics

Four machine learning algorithms are used for the analysis, including logistic regression, support vector machine(SVM), random forest, and multilayer perceptron(MLPs). GridSearchCV and RandomizedSearchCV were applied for tuning the parameters.

Logistic regression is chosen for its effectiveness in binary classification problems, like predicting stroke risk (yes=1, no=0). It's particularly well-suited for datasets with linear relationships, which might be present in attributes like age, hypertension, and heart disease. This will be later discussed in Section 5.1. Also, a key advantage of this algorithm is its interpretability, as we can have an understanding of how each predictor variable (such as gender, smoking status, and bmi value) influences the probability of having a stroke. Furthermore, logistic regression is also computationally efficient, so it is a well-suited starting point for getting insight and establishing a baseline performance.

Support vector machine (SVM) is chosen for its robustness in handling high-dimensional data and it is well-suited for classification tasks due to its effectiveness. Particularly, it is adept at finding the optimal hyperplane that maximizes the margin between the different classes. Hence, this makes it a well-suited one for distinguishing the patients with and without stroke risk, while considering the complex relationship between variables like average glucose level and bmi values, which are normalized in this case. In addition, the kernel trick of the SVM allows it to handle non-linear relationships, which are potentially present in this dataset.

Random forest is chosen for its high accuracy and robustness, especially in handling datasets with imbalance classes, which is potentially present in this stroke situation. It is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes of each individual tree. It is also effective in capturing non-linear relationships without much extensive data preprocessing, which is well-suited for this data. Additionally, it can provide useful insight into the important feature, demonstrating which predictor variables might significantly contribute to the prediction.

Multilayer perceptron (MLP) is chosen for its ability to capture complex and non-linear relationships in the data, and as being a type of neural network, it has multiple layers and

neurons. This might be the potentially most suited algorithm for this project, because the predictor variables have a mix of categorical, binary, and continuous variables. In addition, as discussed in the logistic regression, there might be interaction between variables like age and marriage status, MLP could handle this. Even though its advantages seem perfect, it still has disadvantages. It requires careful hyperparameter tuning and could cause overfitting without proper regularization.

As mentioned above, four machine learning algorithms are used. Actually, to compare the performance of different algorithms, seven models are trained, including baseline logistic regression model, logistic regression model with a threshold of 0.6, tuned and improved logistic regression model, baseline support vector machine model, tuned and improved support vector machine model, random forest model, and multilayer perceptron model.

2.4. Hyperparameter Tuning

The logistic regression model's tuning involves GridSearchCV, focusing on 'C', a regularization parameter, and 'penalty', determining the type of regularization (L1 or L2). The chosen values for 'C' aimed to balance model complexity and performance. In this situation, by performing GridSearchCV, the ‘C’ is chosen to be 1, and regularization is L2, which is ridge regression that is used to prevent overfitting and may address the issue of multicollinearity to certain extent.

For SVM, RandomizedSearchCV was used, tuning parameters like 'C' (penalty parameter), 'kernel' (specifying the kernel type), and 'gamma' (kernel coefficient).

Random forest's tuning via GridSearchCV involves parameters like 'n_estimators' (number of trees), 'max_depth' (depth of each tree), and 'min_samples_split' (minimum number of samples required to split a node), focusing on optimizing model accuracy and preventing overfitting. In this situation, by performing GridSearchCV, the best parameters found are that bootstrop=false, ‘max_depth’=30, ‘min_samples_split’=2, ‘n_estimators’=300.

The MLP's architecture tuning includes layer configurations and neuron counts, targeting an optimal structure for complex pattern recognition in the dataset

3. Results

3.1. Main Findings

As mentioned in the above Section 2.2, seven models are trained, and they are trained based on the training dataset and tuned from validation dataset, and tested on the testing dataset. The

models' final performance on the validation dataset is summarized in Table 2, and performance on the testing dataset is summarized in Table 3, and both tables are attached in Appendix.

According to Table 2, the baseline logistic regression model achieves moderate accuracy(0.7313) and class=1(actual stroke cases) recall of (0.76) but low precision for predicting strokes. The baseline logistic regression model with threshold of 0.6 has a close performance, as accuracy(0.791) and class=1(actual stroke cases) recall(0.71). Compared to the baseline, the improved and tuned logistic regression greatly increases the recall(0.95) for stroke group, but at the cost of accuracy(0.5577) and precision. The baseline SVM has moderate performance overall, with a balance between recall for both groups, but lowers recall(0.53) for stroke detections compared to the baseline LR model. The improved and tuned SVM does not improve accuracy(0.7123), but dramatically improves the recall(0.84) for the stroke group. The random forest model has the highest accuracy(0.8874), and high recall for(0.93) the no stroke group, but very low recall(0.13) for the stroke group. The MLP model has moderate accuracy(0.7666) but balanced recall for the no stroke group(0.77) and the stroke group(0.66). All models have relatively decent ROC-AUC scores, but SVM and RF have relatively lower ones, and except for these two, the other models also suggest strong discriminative abilities.

For the three logistic regression related models, the baseline model and the tuned model with threshold 0.6 are chosen, because there is a trade-off between accuracy and recall for the stroke group, and the baseline with threshold of 0.6 is abandoned because it does not excel in any area compared to the other two. For the baseline and tuned SVM models, the tuned model is chosen while the baseline is abandoned, because the tuned model has decent improvement and triumph over the baseline. The random forest and MLP models are chosen.

I test the models on the testing dataset. According to Table 3, the Base LR model shows a moderate balance between accuracy (0.7151) and recall for both classes, while recall(0.72) for the no stroke group is close to the recall(0.68) for the stroke group. The Tuned LR with a threshold of 0.6 significantly improves recall(0.88) for the stroke group at the expense of overall accuracy (0.5916). The Tuned SVM shows improved recall(0.82) for the stroke group with decent accuracy (0.6947). The RF model exhibits high accuracy (0.8969) but poor recall(0.12) for the stroke group. The MLP model has fairly balanced recall metrics but lower accuracy (0.7558) and recall(0.68) for the stroke group. The ROC-AUC scores vary but are very close, and all above 0.73, indicating strong discriminative abilities. Plot 1 in Appendix is the Receiver

Operating Characteristic (ROC) curve, and it plots the true positive rate against the false positive rate for the different models. A higher area under the curve (AUC) indicates better model performance. Here, the SVM has the highest AUC, indicating it is most effective at distinguishing between stroke and no-stroke cases. Plot 2 in Appendix is the bar chart that displays accuracy and ROC-AUC scores for the models, and it reveals identical results as our previous observation. The random forest model has the highest accuracy, but its ROC-AUC is not the highest, suggesting that while it is accurate overall, it may not be as effective at distinguishing between classes as the SVM.

3.2. Sensitivity Analysis

I observe that imbalance exists in the binary response variable, which is stroke. The training dataset is extremely unbalanced, due to 137 observations with 'no stroke (=0)' while 3297 with 'stroke (=1)'. In other words, the majority of the patients do not have strokes, while having strokes is the minority group. Addressing this issue, I apply SMOTETomek to the training dataset. With SMOTETomek, while SMOTE is an oversampling method which creates synthetic samples for the minority group, which is 'no stroke (=0)' in this case, Tomek Link is a downsampling method which removes the instances of the majority class 'stroke (=1)' from each Tomek Link to clean overlapping points between opposite classes. In general, SMOTETomek is a hybrid method. After applying the SMOTETomek, the training dataset is enlarged to have 6575 observations with 3288 observations in each outcome. As a result, the training dataset is balanced, and used to train the models.

4. Discussion

4.1. Significance of Results

In the realm of stroke prediction, the efficacy of a model is paramount, as timely and accurate identification of potential stroke can significantly influence the patient's outcome. As a result, other than the accuracy of the model, the recall of the stroke group appears to be very significant.

Based on Table 3 in Appendix and description in Section 3.1(Main Findings), I would recommend the improved and tuned SVM model for the prediction of stroke. The reason for the decision is that the baseline logistic regression model and MLP indeed deliver solid performance across the board with no extreme trade-off but do not excel in any particular area, so they are not recommended in this situation. The random forest has a great accuracy(0.8969), but it has a very

low recall(0.12) for the stroke group. Having high accuracy but failing to identify the patients have no meaning in this situation, so it is not chosen as it does not satisfy the need.

Between the tuned LR model with threshold of 0.6 and tuned SVM model, the recalls of the stroke group for the tuned LR(0.88) is higher than tuned SVM(0.82), but the tuned SVM model has higher recall(0.69) for the no stroke group and higher accuracy(0.6947), compared to the tuned LR model with recall(0.58) and accuracy(0.5916). Even though we value the recall for the stroke group, and the tuned LR model has a higher one, the difference between recalls of them is small, while the trade-off for the other metrics are relatively large. In other words, I do focus on higher recall to save more lives by correctly identifying the majority of actual stroke cases, but I do not want to sacrifice too much on the accuracy or correct prediction of no stroke patient. As a result, I would recommend the tuned SVM model for the prediction, because it has moderate accuracy and very high recall for the stroke group.

In clinical practice, such a model could be invaluable for early intervention strategies. It can help in accurately identifying patients at high risk of stroke, which is essential for prompt medical intervention. As early detection is crucial, the model might be able to significantly reduce the risk of long-term disability of a patient and increase the chances of prevention or recovery. Furthermore, this model can be integrated into the healthcare system for pre-screening and hel[doctors in decision-making by highlighting the cases that may require attention.

4.2. Potential Bias

I identify three potential biases within the dataset. The first is the potential sampling bias. The dataset might not fully represent the whole population due to selective data collection, thus affecting the generalizability of my models. The second is potential bias within the sampling process. The observations are assumed to be self-reported, and subjects might underreport or overreport their situations, thus affecting the accuracy of my models. The third one is about the potential confounding variables. This is the relatively significant one, as some other variables might also be significant but not accounted for, thus affecting the predictions.

4.3. Other Insights

Other than solely working with this dataset, this project demonstrated the effectiveness of different machine learning techniques in dealing with real-world medical data, particularly in predicting stroke occurrences.

5. Appendix

5.1. Details of EDA and Analysis

Before implementing the algorithms, I start with basic EDA and relative analysis. First, I observed the correlation matrix between the variables. For better understanding, I generated a correlation heatmap, which is Plot 3 in Appendix. It visualizes the strength and direction of the relationship between the variables. For Plot 3, the darker red shakes indicate a strong positive correlation, where variables move in the same direction, while darker blue shades indicate a strong negative correlation, where variables move in opposite directions. For example, age appears to have a stronger positive correlation with the occurrence of stroke, which is expected since stroke risk will increase as a person gets older. Also, we do see that some are darker blue shades like smoking status and working type. However, these values are not high enough to be problematic interns of multicollinearity. Furthermore, multicollinearity is still considered and mentioned in the previous sections.

In addition to the correlation matrix, I observe the frequency of the variables to see if there is any imbalance or unfairness within the dataset. I do not observe any particular issues in the predictors. One thing that stands out is the imbalance in the binary response variable, and this issue is also addressed in the previous sections.

5.2. Materials Considered

Why is recall used as a relatively more significant factor to consider?

A high recall percentage means the majority of the actual positive cases are correctly diagnosed. This is significant because we might need to identify as many true cases as possible so that proper treatment can be applied promptly. In an ethical angle, failing to identify a patient who actually is ill is way more severe than falsely identifying a health subject. In other words, we will not want to take the risk of not identifying a patient, as the loss is more bitter.

5.3. Tables and Plots

Table 1 : Data Cleaning and Wrangling
Variable *id* is dropped, since it is not useful in this situation.
Variable *gender* is adjusted to be binary, as *Male*=1 and *Female*=0. The value *'other'* is dropped, as there is only one observation.
Variable *hypertension* is adjusted to be binary, as *Yes*=1 and *No*=0.

Variable *heart_disease* is adjusted to be binary, as *Yes*=1 and *No*=0.
Variable *ever_married* is adjusted to be binary, as *Yes*=1 and *No*=0.
Variable *residence_type* is adjusted to be binary, as *Yes*=1 and *No*=0.
Variable *age* is adjusted to be categorical, as *Kids*(0-18)=0, *Young*(18-45)=1, *Middle_aged*(45-62)=2, and *Senior*(>62)=3.
Variable *work_type* is adjusted to be categorical, as *Never_worked*=0, *Private*=1, *Self_employed*=2, and *Govt_job*=3.
Variable *smoking_status* is adjusted to be categorical, as *Never_smoked*=0, *Formly_smoked*=1, *Smokes*=2, and *Unknown*=3.
Variable *bmi* is a float, which is standardized and between 0 and 1.
Variable *avg_glucose_level* is a float, which is standardized and between 0 and 1.
Response variable *Stroke* is binary, as *Yes*=1 and *No*=0.

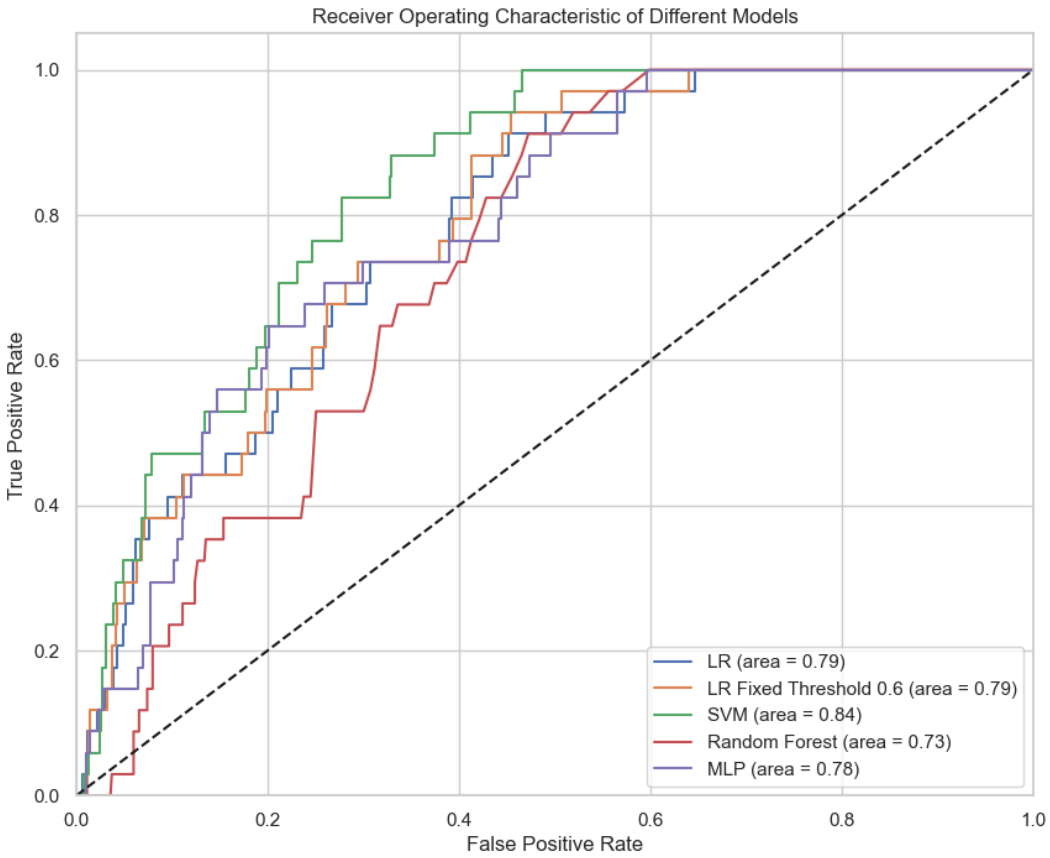
Table 2 - Validation

	Base LR	LR Threshold 0.6	Tuned LR	SVM	Tuned SVM	RF	MLP
accuracy	0.7313	0.7910	0.5577	0.7517	0.7123	0.8874	0.7666
recall(0)	0.73	0.80	0.61	0.76	0.71	0.93	0.77
recall (1)	0.76	0.71	0.95	0.53	0.84	0.13	0.66
precision (0)	0.98	0.98	1.00	0.97	0.99	0.95	0.98
precision (1)	0.13	0.16	0.12	0.11	0.13	0.09	0.14
f1-score (0)	0.84	0.88	0.75	0.85	0.82	0.94	0.86
f1-score (1)	0.23	0.26	0.21	0.18	0.23	0.11	0.23
ROC-AUC	0.8439	0.8439	0.8416	0.7466	0.8478	0.7623	0.8256

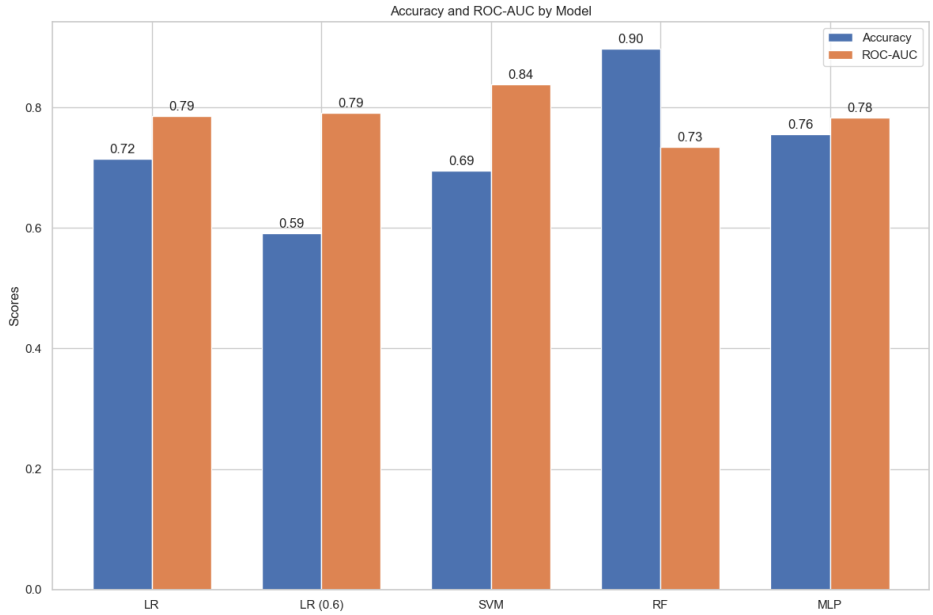
Table 3 - Testing

	Base LR	Tuned LR Threshold 0.6	Tuned SVM	RF	MLP
accuracy	0.7151	0.5916	0.6947	0.8969	0.7558
recall(0)	0.72	0.58	0.69	0.93	0.76
recall (1)	0.68	0.88	0.82	0.12	0.68
precision (0)	0.98	0.99	0.99	0.96	0.98
precision (1)	0.10	0.09	0.11	0.08	0.12
f1-score (0)	0.83	0.73	0.81	0.95	0.86
f1-score (1)	0.18	0.17	0.20	0.10	0.20
ROC-AUC	0.7861	0.7911	0.8383	0.7342	0.7833

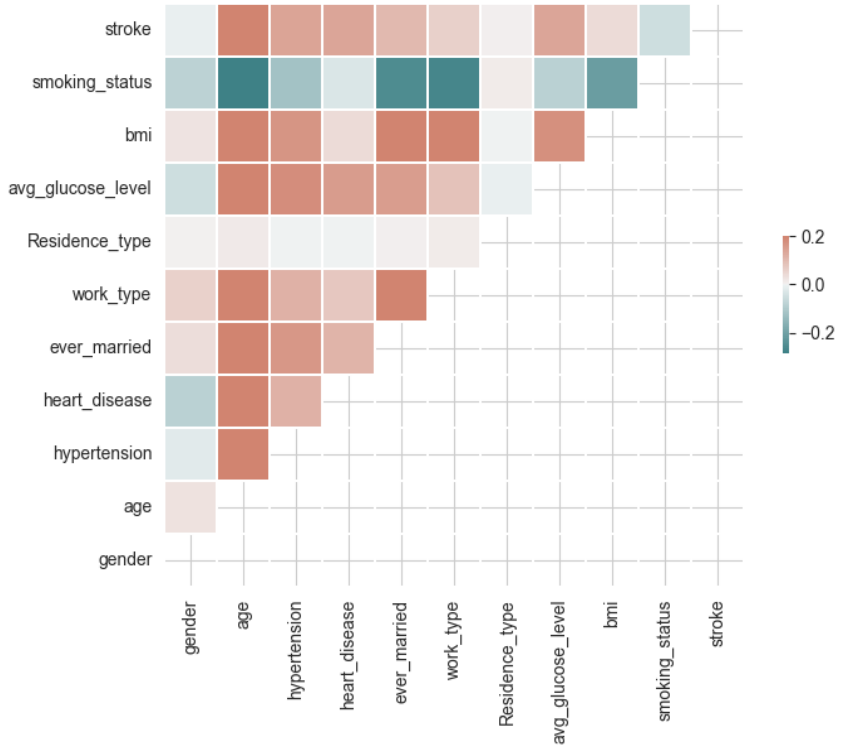
Plot 1



Plot 2


Accuracy and ROC-AUC by Model
Accuracy
ROC-AUC
Scores
0.0
0.2
0.4
0.6
0.8
0.72
0.79
0.59
0.79
0.69
0.84
0.90
0.73
0.76
0.78
LR
LR (0.6)
SVM
RF
MLP

Plot 3


stroke
smoking_status
bmi
avg_glucose_level
Residence_type
work_type
ever_married
heart_disease
hypertension
age
gender
0.2
0.0
−0.2
gender
age
hypertension
heart_disease
ever_married
work_type
Residence_type
avg_glucose_level
bmi
smoking_status
stroke