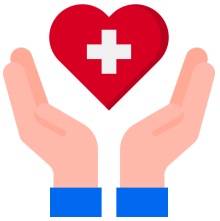
DBA4811 Analytics for Consulting
Academic Year 2024/2025 Semester 2
Final Project Report

Group 2

Stroke Risk Identification for Preventive Healthcare



Name	Matriculation No.
Bryan Teo Jun Hao	[redacted]
Eunice Yap Jia Xin	[redacted]
Jon Fung Zhi Yu	[redacted]
Ng Teng Suan	[redacted]
Teo Hui Jie	[redacted]

Table of Contents

Executive Summary

BrightLife Hospital has observed an increase in stroke-related admissions, prompting the Internal Analytics Team to develop a predictive solution to improve early detection and reduce diagnostic errors. This initiative aligns with the hospital's mission to deliver high-quality, data-driven care while minimizing clinical and reputational risks. Leveraging a dataset of 5,110 patient records with 12 variables, the team conducted thorough data preprocessing, exploratory analysis, and modeling to identify key predictors of stroke. Variables such as age, average glucose level, hypertension, and heart disease emerged as the most significant risk factors.

Several machine learning models were evaluated, including Logistic Regression, Decision Tree, Random Forest, and XGBoost. Logistic Regression—trained exclusively on health-related features—delivered the best overall performance, with a ROC AUC of 0.8518, a recall score of 0.8065, and fast computation time. These results reflect the model's strong ability to detect stroke cases while remaining interpretable and computationally efficient. While models like XGBoost offered competitive accuracy, they lacked the transparency needed for clinical implementation. Random Forest, despite its complexity, performed poorly in recall, making it unsuitable for high-risk medical applications.

To ensure usability in clinical settings, the Logistic Regression model was converted into a point-based scoring system. This Point System assigns simple, interpretable scores to patient factors—such as age brackets and glucose thresholds—making it easy for clinicians to manually calculate stroke risk. For example, a patient aged 65 with elevated glucose, hypertension, and heart disease would score 9 points, corresponding to a 29.9% estimated stroke risk. The system preserves the predictive strength of the original model while enhancing transparency and ease of use.

The model is cost-effective, scalable, and well-suited for deployment across hospital departments, including integration into Electronic Health Record (EHR) systems or mobile applications. It aims to enable more timely interventions, reduce diagnostic errors, and support informed clinical decisions. For continued impact, we recommend piloting the system in real-world clinical workflows, gathering feedback from healthcare professionals, and validating it across broader patient populations.

1. Introduction

BrightLife Hospital is a leading private healthcare provider in Singapore, renowned for its excellence in patient care and medical research. Serving thousands of patients per month, the hospital is committed to delivering high quality healthcare through innovation and data driven decision making.

1.1 Our Service

As key members of BrightLife Hospital's Internal Analytics Team, we harness the power of advanced data analytics and machine learning using Python to drive meaningful insights. Our mission is to enhance patient care, optimize operational efficiency, and address critical challenges faced by the institution. Through data-driven decision-making, we identify strategic opportunities to add value and support the hospital's commitment to excellence in healthcare.

1.2 Business Problem

BrightLife Hospital has observed an increasing trend in stroke-related admissions, which is concerning due to the serious implications of strokes on patient health and hospital operations. According to the World Health Organization (WHO), stroke is among the top three leading causes of death globally. Additionally, it is one of the most commonly misdiagnosed serious conditions in emergency care, with diagnostic errors often resulting in preventable disability or death (Suzuki et al., 2022). The failure to accurately diagnose a stroke can also have severe consequences on the hospital's reputation. Research indicates that patients who experience diagnostic errors frequently lose confidence in their healthcare providers and may avoid seeking care from the same hospital in the future, impacting business performance (Suzuki et al., 2022).

Our team's primary challenge is the early detection and accurate diagnosis of strokes. The hospital recognizes a need to improve its diagnostic capabilities to prevent these issues. The first goal is to more accurately identify at-risk patients early by uncovering patterns that distinguish stroke patients. The second goal is to develop efficient strategies to enhance stroke prevention and reduce the rate of misdiagnosis.

2. Data

2.1 Data Overview

The data contains 5110 observations and 12 attributes. Each row in the data provides relevant information about each patient in the hospital. A description of each attribute is provided in *Appendix 1*. The target variable is the binary variable "stroke".

2.2 Data Preprocessing

We first examined the dataset for missing values and identified 201 missing entries in the "bmi" column. To address this, we replaced the missing values with the median of "bmi". Additionally, categorical variables such as "gender", "ever_married", "work_type", "residence_type" and "smoking_status" were standardized with more descriptive labels to maintain a consistent style. Column names are also reformatted to follow the snake_case convention for consistency in Python. The "id" column was removed as it was completely unique and had no correlation with stroke (*Appendix 2*). Furthermore, categorical variables were transformed using one-hot encoding to convert them into numerical representations, enabling more effective analysis and predictive analytics (*Appendix 3*).

2.3 Exploratory Data Analysis

The dataset showed a significant class imbalance, with just 4.9% of patients diagnosed with stroke and 95.1% without (*Appendix 4*).

For more consistent and interpretable data visualizations, our group divided the dataset into categorical and numeric dataframes (*Appendix 5*) excluding the dependent variable "stroke". For numeric variables, our group generated a histogram, probability plot and boxplot (*Appendix 6*). From these visualizations, the age distribution appears relatively uniform, without a clear skew. The probability plot shows deviations from normality at both tails, indicating that the age distribution does not strictly follow a normal curve. The box plot confirms a well-spread age range with no significant outliers, suggesting a balanced dataset.

The avg_glucose_level distribution is highly right-skewed, with most values clustering between 70 and 150 but extending beyond 250. The probability plot reveals strong deviation from normality especially in the upper range. The boxplot shows numerous high-end outliers, indicating a subset of individuals with abnormally high glucose levels, potentially linked to diabetes or metabolic disorders. Therefore, applying a log transformation may help to normalize the "avg_glucose_level" variable.

The BMI distribution is right-skewed. with most values between 15 and 40 but a long tail extending beyond 60. The probability plot confirms non-normality, as higher BMI values deviate from the expected normal trend. The boxplot highlights multiple extreme outliers above 50, suggesting a subset of individuals with significantly higher BMI.

A scatter matrix is also created to analyse the relationships between numeric features (*Appendix 7*). From this visualization, it suggests that while age and glucose level may be important predictors of stroke risk, no single feature or simple combination of these three features alone creates a clear boundary between the groups, indicating that stroke prediction likely requires more complex modeling approaches or additional features.

For each categorical variable, our group generated a bar chart comparison between stroke and non-stroke patients and obtained the stroke rate of various categorical variables (*Appendix 8*). From these insights, we analysed that hypertension (13.25%), heart disease (17.03%), being married (6.56%) and being a former (7.91%) or current smoker (5.32%) are associated with higher stroke rates in this dataset.

Apart from drawing insights from numeric and categorical analysis, a correlation matrix is also performed to find relationships across all features in Appendix 3. The heatmap reveals strong correlations, such as age with "ever_married_yes" (0.68) and age with "employment_status_unemployment" (-0.64), indicating older individuals are more likely married and employed. "bmi" and "ever_married_yes" (0.33) and "hypertension" and "age" are not strongly related, while "smoker_status_unknown" and "employement_status_unemployement" (0.51) may indicate missing data patterns. These findings help in feature selection by identifying redundancy (e.g., marital status) and ensuring weak predictors don't add noise to models.

The preliminary logistic regression analysis (*Appendix 10*) identified age, hypertension, and average glucose levels as statistically significant predictors of stroke, each with p-values below 0.05. These results are consistent with established medical literature on stroke risk factors. Age demonstrates the strongest influence, with stroke odds increasing by 0.0717 for

every additional year, underscoring its central role. Hypertension also contributes significantly, affirming its status as a key health-related determinant of stroke. Meanwhile, elevated glucose levels are linked to a 0.0041 increase in stroke probability per unit rise—an effect that, while small, remains statistically significant and highlights the critical importance of blood sugar regulation. On the other hand, non-health-related variables, including marital status, employment status, and residential setting (urban vs. rural), were not found to significantly influence stroke risk in this model.

3. Methods

3.1 Data Preprocessing

Prior to model training, we applied further preprocessing steps to prepare the data.

3.1.1 Log Transformation

During exploratory data analysis, we observed that the avg_glucose_level variable was highly right-skewed, with a long tail of abnormally high values. This kind of skewness can negatively impact model performance, especially for algorithms like logistic regression which assume linearity and are sensitive to the scale and distribution of input features. To address this, we applied a logarithmic transformation using np.log, which effectively compresses the range of high glucose values while preserving the order and spacing of smaller values. This transformation helps normalise the distribution, reduce the influence of extreme outliers, and stabilise variance—resulting in a more balanced and model-friendly feature.

3.1.2 Train-Test Split

To ensure a fair and unbiased evaluation of model performance, we split the dataset into training and test sets before performing any data preprocessing steps—this includes operations such as log transformation and feature scaling. Conducting these steps after the train-test split is a critical practice as it prevents data leakage, a situation where information from the test set unintentionally influences the model during training. Data leakage can lead to overly optimistic performance metrics and models that fail to generalise to unseen data. By keeping the test set completely isolated during training, we ensured that our evaluation metrics would accurately reflect the model's real-world performance on new patients.

3.1.3 Standard Scaling

To prepare the dataset for model training, we applied standard scaling to our continuous variables. Standard scaling, also known as z-score normalisation, transforms each feature so that it has a mean of zero and a standard deviation of one. This process ensures that all continuous variables are on the same scale and contribute equally to the model during training.

This step is particularly important for models such as logistic regression which are sensitive to the magnitude of input features. Without scaling, features with larger numerical ranges—such as glucose levels—may disproportionately influence the model's learning process compared to features like age or BMI. Standardisation also improves the efficiency and reliability of the model’s training, especially when regularisation techniques are used, as it helps the optimisation algorithm converge more quickly and accurately.

To maintain the integrity of the model evaluation process and avoid data leakage, we applied standard scaling only after splitting the data into training and testing sets. The scaler was fitted exclusively on the training set, and the same transformation was applied to the test set

using the parameters learned from the training data. This ensures that no information from the test set is used during model training, resulting in a more reliable assessment of the model's performance.

3.2 Cross-Validation Techniques

In building a clinically relevant stroke prediction model, it was important not only to train a model that performs well statistically, but also one that reflects the real-world consequences of misclassification in a healthcare setting. To achieve this, we employed a two-stage approach to optimize our model's performance: hyperparameter tuning and classification threshold adjustment. This allowed us to refine both the internal structure of the model and its decision-making criteria to better suit the clinical context of stroke prediction.

3.2.1 Data Balancing

Before any model tuning, we first addressed the class imbalance problem in our dataset, where only 4.9% of patients had experienced a stroke. Imbalanced datasets can cause models to become biased toward the majority class, leading to poor detection of rare but critical cases—in our case, patients at risk of stroke.

While synthetic oversampling techniques like SMOTE are commonly used to increase representation of minority classes, we chose not to apply it here. In clinical datasets, artificially generating new examples based on feature interpolation may distort important medical patterns or violate physiological plausibility (Alkhawaldeh, I. M. et al., 2023). Preserving the authenticity of patient data was crucial for maintaining clinical relevance.

Instead, we adopted a model-integrated balancing strategy. For all models except XGBoost, we applied class_weight="balanced", which automatically adjusts class weights inversely proportional to class frequencies. This ensures that both classes contribute equally to the model's learning process, without altering the original data distribution. When paired with regularisation, this approach helps models treat minority cases more equitably and reduces bias toward the majority class.

For XGBoost, which does not support class_weight, we used the equivalent scale_pos_weight=19—a value derived from the ratio of negative to positive cases in the training set. This adjustment penalises false negatives more heavily during training, aligning with our clinical priority of ensuring stroke cases are correctly identified.

3.2.2 Cross-Validation and Threshold Optimisation

Once the data was properly prepared and class imbalance accounted for, we turned our attention to model tuning and evaluation. We began by applying GridSearchCV, a technique that performs an exhaustive search across combinations of model hyperparameters using k-fold cross-validation. This approach further ensures we avoid overfitting to a specific subset of patients.

For this task, we prioritised recall as our main scoring metric—it is crucial in clinical settings because it reflects the model’s ability to correctly identify actual stroke cases out of all patients diagnosed with stroke. False negatives—patients who are incorrectly predicted as low-risk—can lead to missed diagnoses, delayed interventions, and serious complications. GridSearchCV was applied across the range of classifiers we used which included logistic regression with regularisation, decision trees, random forests, and XGBoost. This comprehensive approach enabled a robust comparison of models based on recall performance.

After identifying the best-performing model, we performed an additional refinement using TunedThresholdClassifierCV. Most classification models default to a threshold of 0.5 for converting predicted probabilities into class labels. However, this threshold may not be optimal for imbalanced datasets. TunedThresholdClassifierCV searches for a probability cutoff that maximizes a chosen metric, which in our case was balanced accuracy.

Balanced accuracy considers both true positive and true negative rates equally, making it a more appropriate measure when both types of misclassification—false positives and false negatives—have real-world consequences. In the healthcare context, this means improving both the detection of stroke patients and the avoidance of unnecessary concern for low-risk individuals. Additionally, while our earlier GridSearchCV optimization prioritised recall (to minimize false negatives), tuning for balanced accuracy introduces an opportunity to recover some lost precision—resulting in a more clinically practical trade-off.

By maximizing balanced accuracy, we also shift the model's position further toward the top-left of the ROC curve, indicating an overall improvement in true positive and/or true negative rates. Importantly, this threshold tuning makes the model more flexible: hospitals can later adjust the threshold to emphasize either sensitivity or specificity based on institutional needs. For example, a facility with limited capacity might prioritize true negatives to reduce unnecessary screenings, while another focused on early intervention may emphasize true positives. Our model was designed to be both readable and adaptable, empowering practitioners to make threshold-based decisions grounded in their operational context.

By combining hyperparameter tuning (via GridSearchCV) and threshold optimization (via TunedThresholdClassifierCV), we developed a stroke prediction model that balances clinical relevance with statistical rigor. This ensures that model outputs are not only mathematically sound but also practically useful for clinicians making life-impacting decisions.

3.4 Models

We chose to use the following models in our training phase: Logistic Regression, Decision Tree Classifier, Random Forest Classifier and XGBoost Classifier, then compare their performances to determine the model best suited for this business use case (Appendix 11).
For the modelling, we did 2 different approaches. The first approach is using all available features, followed by the second approach that focuses solely on health-related features, excluding marriage status, residence type and employment status.

This comparison was carried out to evaluate the impact of including non-health variables on model performance and to determine whether a more targeted, health-only feature set could achieve similar predictive power with improved interpretability given our knowledge of the medical context.

To ensure reproducibility and maintain consistent results across different runs of the model, we chose to define certain fixed parameters. A random state value of 42 was set to produce the same train-test splits and model behavior. For model evaluation, the number of cross-validation splits was fixed at 5. The proportion of data reserved for testing (test size) was fixed at 20%. Lastly, we chose the following numeric features ("age", "avg_glucose_level", "bmi") so that scaling could be applied only after the train-test split.

3.4.1 Logistic Regression

We developed a logistic regression model using ElasticNet regularization (which combines L1 and L2 penalties) to help prioritize the most important features and prevent overfitting. We used GridSearchCV with the parameters C=0.01 and l1_ratio=0.0 for both approaches.

For Approach 1, where all features are included, the model performed well with a test ROC AUC score of 0.8516, indicating strong overall classification ability. To further improve performance, we adjusted the decision threshold from the default 0.5 to 0.5359, which helped increase the recall to 0.7581—meaning it was better at identifying actual stroke cases—while maintaining an accuracy of 0.7701. For Approach 2, where only health-related features are included, the model still achieved very similar results with a test ROC AUC score of 0.8518, recall: 0.7581, accuracy: 0.774, proving that these non-health features contributed little to prediction quality.

With a slightly adjusted threshold of 0.5434, we again improved recall while keeping model performance stable. Also, training time remained low at under 0.04 seconds, showing that the model is both effective and efficient.

3.4.2 Decision Tree

We built a Decision Tree model for stroke prediction and tuned it to limit complexity and avoid overfitting by setting a maximum depth of 5 for both approaches.

For Approach 1, where all features are included, the model showed a moderate recall of 0.6452, meaning it missed a fair number of actual stroke cases. After adjusting the prediction threshold to 0.6869, recall did not improve where it remained at 0.6452, although the F1 score rose slightly to 0.2996, indicating a small gain in balancing precision and recall. Accuracy stayed steady at 0.817, and the model remained fast to train.

For Approach 2, where only health-related features are included, performance stayed the same where the recall score is at 0.6452, accuracy at 0.817, and a similar improvement in F1 score after tuning to 0.2996. This consistency suggests that removing non-health features did not harm predictive ability, which may support a more privacy-conscious approach without losing accuracy.

3.4.3 Random Forest Classifier

We used a Random Forest model to predict stroke risk, setting the number of trees to 300 and limiting their depth to 10 to manage complexity for both approaches.

For Approach 1, where all features are included, the model delivered strong performance, with a high ROC AUC of 0.8473, meaning it was good at distinguishing between stroke and non-stroke cases. However, after adjusting the decision threshold to favor recall where it increases to 0.8548, accuracy dropped to 0.7417, and the F1 score rose slightly 0.2865, reflecting a better balance between catching true cases and avoiding false alarms. One drawback was that the average training time was around 5.35 seconds.
For Approach 2, where only health-related features are included. Surprisingly, removing non-health variables led to a significant jump in recall to 0.9032, meaning the model caught more actual stroke cases. This suggests that non-health data may have introduced noise rather than added value. However, this came at a cost where the accuracy dropped to 0.6458, meaning more incorrect predictions overall.

3.4.4 XGBoost Classifier

The XGBoost model is a boosted tree approach that works by correcting mistakes made by earlier trees.

For Approach 1, where all features are included and with settings of max_depth=3 and n_estimators=100, it showed strong overall performance. This model achieved a high ROC AUC of 0.8410, meaning it could effectively distinguish between stroke and non-stroke cases. Recall was also solid at 0.7419, suggesting the model was good at catching actual stroke cases. After adjusting the decision threshold to fine-tune its predictions, accuracy reached 0.774 and the F1 score was 0.2848, reflecting a good balance between precision and recall. The model also trained fairly quickly under 0.2 seconds, making it efficient.

For Approach 2, where only health-related features are included. The model still performed well. Recall remained high at 0.7258, while accuracy slightly improved to 0.7710. This suggests that excluding unrelated features helped reduce noise, making the predictions more reliable. The F1 score after tuning remained similar at 0.2778, showing consistent performance even with fewer inputs.

4. Evaluation

4.1 Confusion Matrix

The confusion matrix shows how well the model predicts each class, in our case, whether a patient has had a stroke or not. It breaks the predictions into four categories:

- True Positives (TP): correctly identified stroke cases
- True Negatives (TN): correctly identified non-stroke cases
- False Positives (FP): predicted stroke when there wasn't one (false alarm)
- False Negatives (FN): missed actual stroke cases

In a healthcare setting like this, both false negatives and false positives can have serious consequences. So looking at the confusion matrix helps ensure we are understanding the trade-offs.

Logistic Regression Model: Approach 1 (Appendix 12), Approach 2 (Appendix 13)

After threshold tuning, both approaches show identical performance in stroke detection, but Approach 2 demonstrates a slight edge in precision and specificity due to fewer false positives.

- True Positives (TP): Both approaches correctly identified 47 stroke cases after threshold tuning — no difference observed.
- False Negatives (FN): Missed stroke cases are the same in both approaches, with 15 false negatives each — recall remains unchanged.
- False Positives (FP): Approach 1 produced 220 false positives, while Approach 2 had slightly fewer at 216 — indicating marginally better precision in Approach 2.
- True Negatives (TN): Approach 1 correctly classified 740 non-stroke cases, while Approach 2 classified 744 correctly — suggesting slightly better specificity in Approach 2.

Decision Tree Classification Model: Approach 1 (Appendix 14), Approach 2 (Appendix 15)
After threshold tuning, both Decision Tree approaches perform identically across all classification metrics.

- True Positives (TP): Both approaches correctly identified 40 stroke cases — no difference observed.
- False Negatives (FN): Each model missed 22 stroke cases — recall remains the same.
- False Positives (FP): Both approaches yielded 165 false positives — no gain or loss in precision.
- True Negatives (TN): Both correctly classified 795 non-stroke cases — specificity is equal.

Random Forest Classifier Model: Approach 1 (Appendix 16), Approach 2 (Appendix 17)
After threshold tuning, both approaches improve stroke detection, but Approach 2 outperforms slightly in recall, while Approach 1 retains higher precision and specificity.

- True Positives (TP): Approach 1 correctly identified 53 stroke cases, while Approach 2 identified slightly more at 56 — showing better recall in Approach 2.
- False Negatives (FN): Missed stroke cases were lower in Approach 2 (6 vs. 9), further confirming improved sensitivity.
- False Positives (FP): Approach 1 produced 255 false positives, while Approach 2 yielded more at 356 — better precision in Approach 1.
- True Negatives (TN): Approach 1 correctly classified 705 non-stroke cases, while Approach 2 managed only 604 — higher specificity in Approach 1.

XGBoost Classifier Model: Approach 1 (Appendix 18), Approach 2 (Appendix 19)
Both XGBoost approaches show nearly identical results after threshold tuning, with Approach 1 having a very slight edge in recall and specificity.

- True Positives (TP): Approach 1 correctly identified 46 stroke cases, while Approach 2 identified 45 — a negligible advantage for Approach 1.
- False Negatives (FN): Approach 1 missed 16 stroke cases; Approach 2 missed 17 — recall is marginally better in Approach 1.
- False Positives (FP): Approach 1 had 215 false positives versus 217 in Approach 2 — slightly better precision in Approach 1.
- True Negatives (TN): Approach 1 correctly classified 745 non-stroke cases, while Approach 2 had 743 — again, a tiny edge for Approach 1.

Overall, **Logistic Regression (Approach 2)** is the most well-rounded and dependable model, especially in medical contexts where minimizing both false negatives (missed strokes) and false positives (unnecessary alerts) is essential.

4.2 Recall Score

The recall score (that can be found in Appendix 11), which measures a model's ability to correctly identify actual positive cases, is a critical metric for healthcare-related predictions where false negatives can be costly.

For Approach 1: Among the evaluated models, Logistic Regression achieved the highest recall score of 0.8065, meaning it correctly identified approximately 81% of stroke cases. This makes it the most reliable model when the goal is to minimize missed diagnoses.

XGBoost followed closely with a recall score of 0.7419, still capturing a strong proportion of true stroke cases and offering a good balance between sensitivity and other metrics. The Decision Tree Classifier showed a moderate recall of 0.6452, indicating a fair performance but missing more stroke cases compared to the top two models. In contrast, the Random Forest Classifier had the lowest recall at 0.3065, identifying less than a third of actual stroke cases, which makes it the least suitable option when high sensitivity is essential.

For Approach 2: Logistic Regression again demonstrated the highest recall score of 0.8065, effectively identifying around 81% of true stroke cases. This confirms its strength as it consistently performs well despite its simplicity, which is especially important in medical applications where failing to detect a positive case (stroke) could have serious consequences. XGBoost followed with a recall of 0.7258, showing it is also highly effective at capturing positive cases while potentially offering better performance on other metrics like AUC. The Decision Tree Classifier posted a moderate recall of 0.6452, still reasonable but less reliable than the top two for minimizing false negatives. Notably, the Random Forest Classifier had the lowest recall of 0.2581, identifying only about 26% of actual stroke cases, which severely limits its usefulness when recall is a top priority.

Overall, **Logistic Regression** continues to be more consistent across both approaches, making it the most appropriate model where early and accurate detection is critical.

4.3 ROC AUC score

The ROC AUC score helps us assess the model's ability to distinguish between stroke and non-stroke cases, regardless of the decision threshold. A score of 1.0 means perfect separation, while a score of 0.5 means the model is no better than random guessing. The ROC AUC gives us a big-picture view of how good the model is overall, not just at one threshold, but across many possible thresholds.

In evaluating all models, we considered ROC AUC score, mean fit time, the effect of threshold tuning on confusion matrix performance, and the tradeoffs between precision and recall. Logistic Regression achieved the highest ROC AUC in both approaches (0.8516 in Approach 1 and 0.8518 in Approach 2), indicating excellent ability to discriminate between stroke and non-stroke cases and suggesting low risk of overfitting. XGBoost also performed well with AUCs above 0.84, followed by Random Forest, while Decision Tree scored lowest at 0.7946. In terms of mean fit time, Logistic Regression and Decision Tree were the fastest, both under 0.05 seconds. XGBoost was moderate, while Random Forest was the slowest, taking up to 5.3 seconds in Approach 1. The mean fit time measures how long, on average, it takes for a model to train (or "fit") on the data, so a faster model is also preferred.

Threshold tuning did not affect AUC scores but significantly influenced precision and recall. Logistic Regression (Approach 2) maintained the best overall balance — correctly identifying 47 stroke cases (true positives) while limiting false positives to 216 and keeping false negatives at 15 (Appendix 13). Random Forest (Approach 2) achieved slightly higher recall with 56 true positives and only 6 false negatives, but at the cost of 356 false positives, making it less precise (Appendix 15). XGBoost and Decision Tree showed smaller performance gains and remained consistent between approaches (Appendix 17 and 19).

Given the business context of stroke prediction, where missing cases can have critical consequences but too many false alarms can also overwhelm clinicians, **Logistic Regression (Approach 2)** offers the most balanced and reliable performance. It combines strong AUC,

high recall, low false positive rate, and fast fit time, making it the most suitable model for implementation as a real-time clinical scoring tool.

5. End User Implementation (Points System)

Although models like XGBoost and Random Forest achieved higher recall and AUC scores, they are less interpretable and require longer computation times. In contrast, Logistic Regression and Decision Trees demonstrate lower recall and AUC but offer greater interpretability and faster computation, making them more suitable for clinical settings.

Benefits & Importance of Points Systems

Providing a Logistic Regression equation may be daunting for non-experts to use. Therefore, the purpose of implementing a Point System is to help stakeholders such as physicians and patients at BrightLife Hospital calculate the risk of stroke more easily. Building this Point System enhances transparency by translating complex model coefficients into an intuitive scoring framework, enabling stakeholders to understand and manually estimate stroke risk using a straightforward additive scorecard. This gives insights to patients on what they can do to lower risk and how much it could be lowered, thus making the model both accessible and clinically practical.

Model & Feature Selection

A Point System is built using the best-performing logistic regression model trained exclusively on health-related features (Approach 2). This approach was intentional to maintain interpretability and ensure the inclusion of medically meaningful predictors. From this model, we retained only the features that demonstrated substantial impact on stroke prediction, defined as having logistic regression coefficients greater than 0.20. These features included Age, Average Glucose Level, Hypertension, and Heart Disease. Variables that exhibited minimal predictive power or introduced ambiguity were excluded. For instance, attributes such as “smoker”, “never” and “unknown” smoking status were omitted due to their vagueness, which could reduce clarity and introduce noise into the scoring framework.

Interpreting Coefficients

To make the model clinically relevant and easy to understand, we reverse-transformed the input data used in training. Standardized variables like age and average glucose level were returned to their original scales, such as years and milligrams per deciliter (mg/dL), respectively. Additionally, any log-transformed features were exponentiated. We retained only the four most impactful variables identified earlier, ensuring the model remained focused and interpretable. This transformation allowed us to present the logistic regression coefficients in real-world units, making the scoring logic more relatable and actionable for clinicians and patients alike.

A critical step in building the point system was translating the raw logistic regression coefficients into a scoring scale that was both fair and easy to use. This required the use of custom scaling factors, rather than a single fixed divisor. The reason for this is that logistic regression coefficients naturally exist on very different scales. For example, the coefficient for age was significantly larger than those for hypertension or heart disease.

Using a single scaling factor, such as dividing all coefficients by 0.3, would have resulted in some variables—especially binary ones like hypertension and heart disease—contributing less than one point to the score. This would effectively reduce their influence to zero, misrepresenting their clinical importance. To prevent this, we applied tailored scaling: age

was assigned a larger divisor to scale in 5-year intervals, preventing it from overshadowing other features; glucose was scaled to yield roughly 1 point for every 25 mg/dL increase; and binary variables were given a consistent base of approximately 1 point per positive case. This approach ensured that all variables contribute meaningfully, reflected their statistical weight, and preserved clinical interpretability.

Building the Points System

Variable	Category	Reference Value	Base Difference	Logit Units (Coefficient × Base Diff)	Points Calculation (Logit Units / Scaling Factor)	Points
Age	<30	25	25 - 37 = -12	-12 × 1.284613 = -15.415356	-15.415356 ÷ (5×1.284613) = -2.4 → **-2**	-2
	30-44	37	0	0	0 ÷ 6.423065 = 0 → **0**	0
	45-60	52.5	+15.5	+15.5 × 1.284613 = +19.911502	+19.911502 ÷ 6.423065 = +3.1 → **+3**	+3
	61-70	65.5	+28.5	+28.5 × 1.284613 = +36.611470	+36.611470 ÷ 6.423065 = +5.7 → **+6**	+6
	≥71	76.5	+39.5	+39.5 × 1.284613 = +50.742214	+50.742214 ÷ 6.423065 = +7.9 → **+8**	+8
Avg Glucose mg/dL	<80	70	-20	-20 × 0.212436 = -4.248720	-4.248720 ÷ (25×0.212436) = -0.8 → **-1**	-1
	80-99	90	0	0	0 ÷ 5.3109 = 0 → **0**	0
	100-125	112.5	+22.5	+22.5 × 0.212436 = +4.779810	+4.779810 ÷ 5.3109 = +0.9 → **+1**	+1
	126-199	162.5	+72.5	+72.5 × 0.212436 = +15.401610	+15.401610 ÷ 5.3109 = +2.9 → **+3**	+3
	≥200	235	+145	+145 × 0.212436 = +30.803220	+30.803220 ÷ 5.3109 = +5.8 → **+6**	+6
Hypertension	No	-	-	0	0 ÷ 0.3 = 0 → **0**	0
	Yes	-	-	0.295013	0.295013 ÷ 0.3 = 0.983 → **+1**	+1
Heart Disease	No	-	-	0	0 ÷ 0.3 = 0 → **0**	0
	Yes	-	-	0.206148	0.206148 ÷ 0.3 = 0.687 → **+1**	+1

Figure 1

Each selected feature was then binned into clinically meaningful categories to facilitate the construction of the scoring system as seen in Figure 1. Age was stratified into groups that align with common stroke risk brackets, such as under 30, 30-44, 45–54, and 55–64 years. Average glucose level was categorized using established clinical thresholds, including the prediabetes cutoff at 100 mg/dL and diabetic cutoff at 126 mg/dL (Diabetes SG, 2022). Binary features like hypertension and heart disease were treated as either present (Yes) or absent (No). Each bin was assigned a point value based on its deviation from a healthy reference category, using a logit-based transformation to maintain the additive structure of logistic regression. This methodology allowed the score to be computed easily and interpreted clearly in a clinical context.

Mapping Points to Risk

Points	Stroke Risk
-3	0.64%
-2	0.90%
-1	1.27%
0	1.80%
1	2.53%
2	3.56%
3	4.97%
4	6.91%
5	9.53%
6	13.01%
7	17.51%
8	23.15%
9	29.94%
10	37.75%
11	46.26%
12	54.98%
13	63.41%
14	71.09%
15	77.73%
16	83.20%

Figure 2

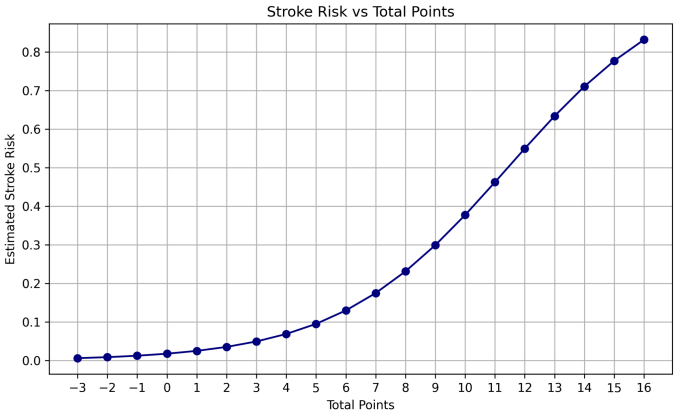


Figure 3

Once the total points have been calculated based on the relevant health features, the overall score is converted into an estimated stroke probability using a lookup table derived from the logistic function as seen in Figure 2. This function produced an S-shaped curve as seen in Figure 3 that aligns total score with estimated stroke risk. Lower total scores were associated with low probabilities of stroke, while higher scores indicated substantially elevated risk. To avoid overconfidence in the model's predictions, we capped the score at 16 points, which corresponded to a stroke risk of approximately 83.2%. This upper limit ensures that the system remains practical and actionable, without creating unnecessary alarm in cases where the probability nears certainty.

Application

For example, consider a 65-year-old patient with an average glucose level of 122 mg/dL who has both hypertension and heart disease. To estimate their stroke risk, the patient can refer to Figure 1 to calculate their total score:

- Age (65 years): +6 points
- Average Glucose (122 mg/dL): +1 point
- Hypertension: +1 point
- Heart Disease: +1 point

This results in a total score of 9 points. By referencing Figure 2, the patient can then determine their estimated stroke risk, which corresponds to a 29.9% likelihood of experiencing a stroke.

The final scoring system in Figure 3 is both clinically grounded and highly interpretable. It preserves the direction and relative magnitude of the original logistic regression model while avoiding the overemphasis or underrepresentation of any individual feature. The score generates a smooth and monotonic risk curve ranging from 0 to 16 points, allowing for straightforward estimation of stroke probability.

This system works effectively for several reasons. First, custom scaling factors were used to ensure each variable contributed appropriately to the score. Second, point values were aligned with real-world clinical thresholds, making them intuitive for medical professionals. Third, the overall design strikes a balance between simplicity and predictive performance. Finally, the system is ready for implementation in clinical tools such as electronic health record (EHR) decision aids, mobile applications, or point-of-care dashboards, enhancing its utility in day-to-day healthcare settings.

6. Conclusion

The client's request aligns closely with BrightLife Hospital's core mission of clinical excellence and innovation. By enabling early stroke detection, the project contributes to reduced patient morbidity, lower healthcare costs, and increased patient confidence—delivering strong value both operationally and reputationally. Moving forward, involving hospital leadership during the early planning stages could help ensure the model's objectives are even more closely aligned with the hospital's broader strategic priorities.

We applied strong analytical practices throughout the project—conducting thorough exploratory data analysis (EDA), addressing skewed distributions through log transformation, preventing data leakage by applying preprocessing after the train-test split, and fine-tuning thresholds to manage class imbalance effectively. While class weighting addressed imbalance appropriately, future iterations could explore ensemble techniques such as bagging or hybrid models to enhance robustness, particularly when generalizing to external datasets.

Additionally, the high AUC observed may indicate potential overfitting, suggesting that further regularization may be necessary to improve model generalizability.

The analytics successfully achieved the primary goal of predicting stroke risk and extended beyond that by delivering a clinically practical solution through the development of a Point System. We strategically shifted from high-performance but complex models like XGBoost to more interpretable models such as Logistic Regression, emphasizing usability and transparency for clinical adoption. To enhance real-world applicability, future iterations should undergo pilot testing within actual clinical workflows to ensure smooth integration with electronic health record (EHR) systems and routine physician practices. Additionally, the Point System may require periodic updates, and external validation across diverse patient populations will be essential to strengthen the model's reliability and generalizability.

The findings are well-supported by statistical analysis and the key predictors identified are consistent with known stroke risk factors. The introduction of the Point System translates complex model outputs into a clear, actionable format, enhancing clinician engagement and understanding. To further strengthen the credibility and relevance of the solution, future work could include benchmarking against existing stroke risk calculators and incorporating feedback from healthcare professionals through interviews or surveys to validate its practical utility.

The proposed solution is both cost-effective and scalable, making it suitable for deployment across various hospital departments. It supports early intervention, minimizes false alarms, and holds significant potential to improve patient outcomes while reinforcing the hospital's reputation. From a business perspective, the solution offers a strong return on investment. However, the lack of real-world validation currently limits clinical confidence. Future iterations should consider controlled trials or post-deployment evaluations to better assess its impact in practice.

The project was communicated with clarity, supported by well-organized appendices and comprehensive documentation. The rationale behind data preprocessing, model selection, and the design of the Point System was clearly articulated and understandable for both technical and non-technical stakeholders. To further enhance communication—particularly at the executive level—future deliverables could benefit from the inclusion of interactive visual aids such as a summary dashboard, point-based scorecard visuals.

References

World Health Organization. (2024, August 7). *The top 10 causes of death*. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Suzuki, R., Yajima, N., Sakurai, K., Oguro, N., Wakita, T., Thom, D. H., & Kurita, N. (2022). Association of Patients' Past Misdiagnosis Experiences with Trust in Their Current Physician Among Japanese Adults. Journal of general internal medicine, 37(5), 1115–1121. https://doi.org/10.1007/s11606-021-06950-y

Alkhawaldeh, I. M., Albalkhi, I., & Naswhan, A. J. (2023, December 20). Challenges and limitations of synthetic minority oversampling techniques in machine learning. World journal of methodology. https://pmc.ncbi.nlm.nih.gov/articles/PMC10789107/

Managing diabetes: Blood glucose. Diabetes SG. (2022, January 12). https://www.diabetes.org.sg/resource/managing-diabetes/

Appendices

Appendix 1: Original Data Overview and Description

Attribute	Description
id	Unique Identifier
gender	"Male", "Female" or "Other"
age	Age of the patient
hypertension	(0 = patient doesn't have hypertension, 1 = patient has hypertension)
heart_disease	(0 = patient doesn't have any heart diseases, 1 = patient has a heart disease)
ever_married	"No" or "Yes"
work_type	"children", "Govt_job", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
avg_glucose_level	Average glucose level of blood (in mg/dL)
bmi	Patients' body mass index
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"
stroke	(1 = patient had a stroke, 0 = patient did not have a stroke)

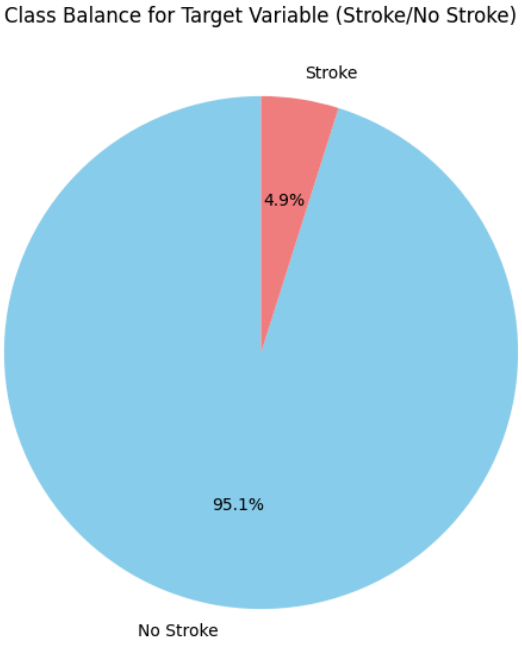
Appendix 2: Renamed Column Names and Categorical Values

Original Attribute	Renamed Attribute	Description
gender	gender	“male” or “female”
age	age	Age of the patient
hypertension	hypertension	(0 = patient doesn't have hypertension, 1 = patient has hypertension)
heart_disease	heart_disease	(0 = patient doesn't have any heart diseases, 1 = patient has a heart disease)
ever_married	ever_married	“no” or “yes”
work_type	employment_status	“Private” is changed to “employed” “Self-employed” is changed to “employed” “children” is changed to “unemployed” “Govt_job” is changed to “employed” “Never_worked” is changed to “unemployed”
Residence_type	residence_type	“rural” or “urban”
avg_glucose_level	avg_glucose_level	Average glucose level in blood
bmi	bmi	Patients’ body mass index
smoking_status	smoker_status	“never smoked” is changed to “never” “Unknown” is changed to “unknown” "formerly smoked" is changed to “former” “smokes” is changed to “smoker”
stroke	stroke	(1 = patient had a stroke, 0 = patient did not have a stroke)

Appendix 3: Perform One-Hot Encoding on New Dataset

Original Attribute	Description
age	Existing Features
avg_glucose_level	
bmi	
stroke	
hypertension	
heart_disease	
gender_male	One-Hot Encoding
ever_married_yes	
employment_status_unemployed	
residence_type_urban	
smoker_status_never	
smoker_status_smoker	
smoker_status_unknown	

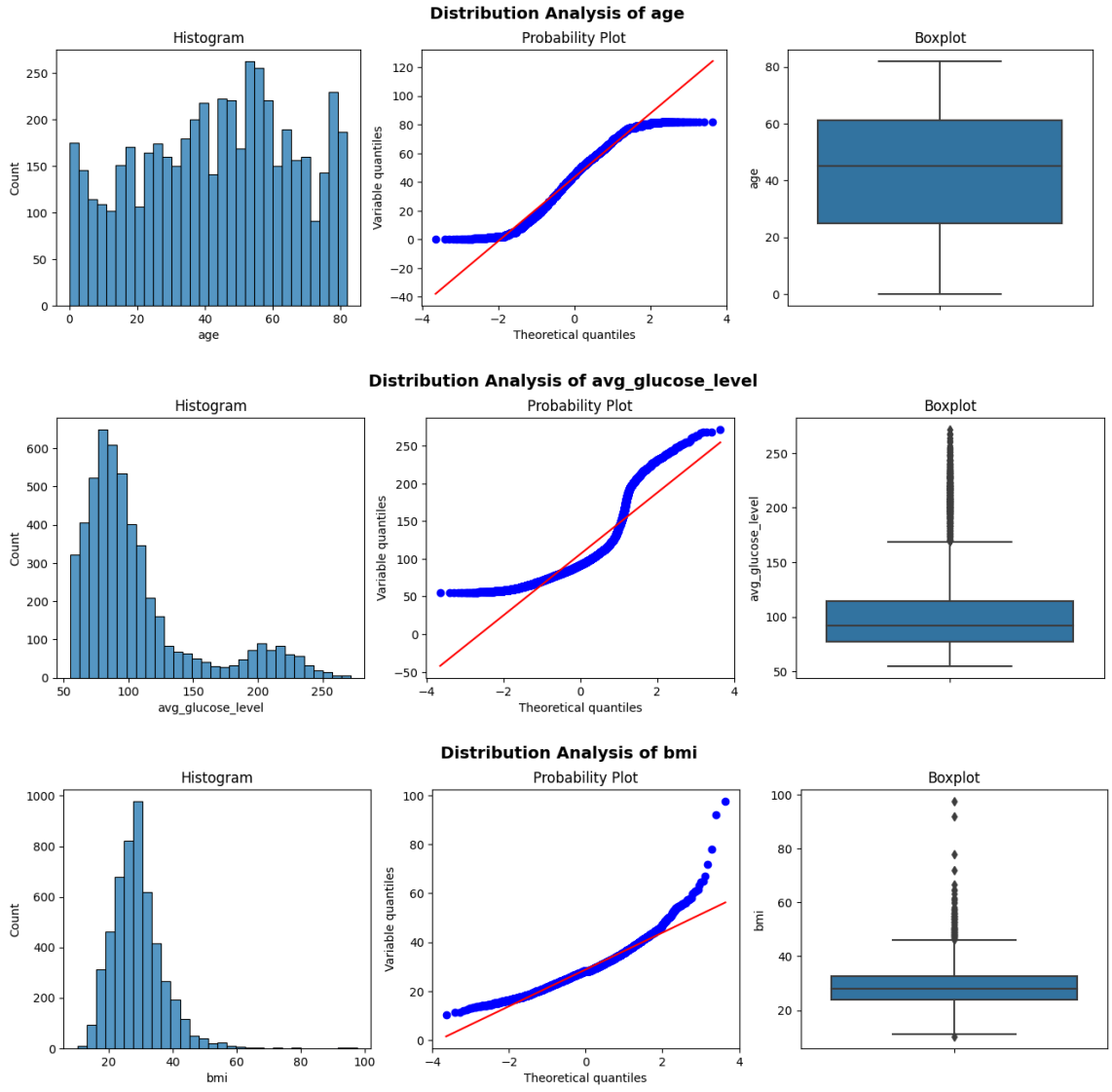
Appendix 4: Proportion of Patients with Stroke



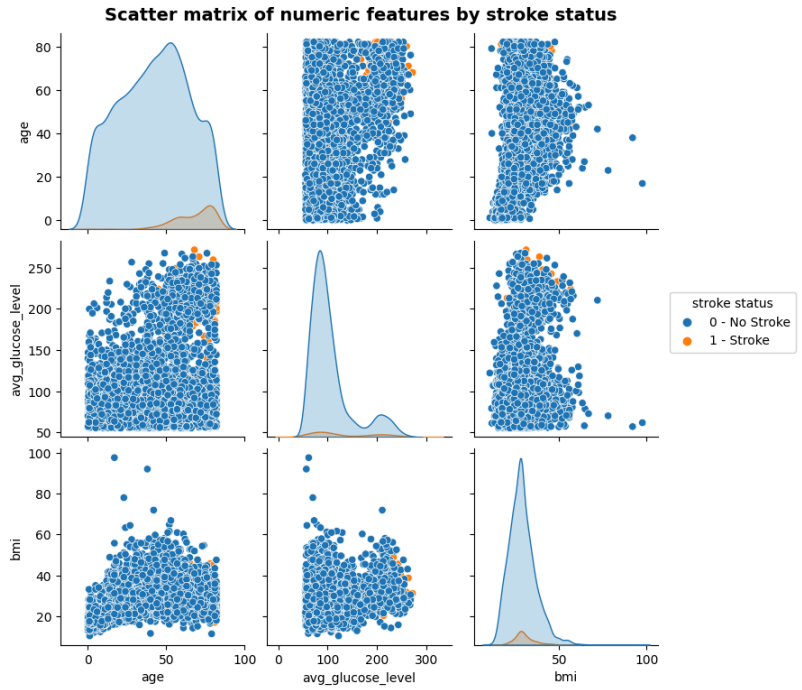
Appendix 5: Numeric Features and Categorical Features in New Dataset

Numeric Features	Categorical Features
age	gender
avg_glucose_level	hypertension
bmi	heart_disease
	ever_married
	employment_status
	residence_type
	smoker_status

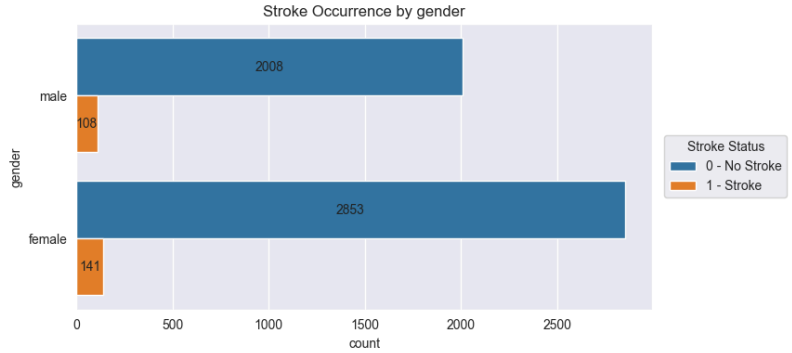
Appendix 6: Numeric Features Analysis



Appendix 7: Numeric Features Scatter Matrix



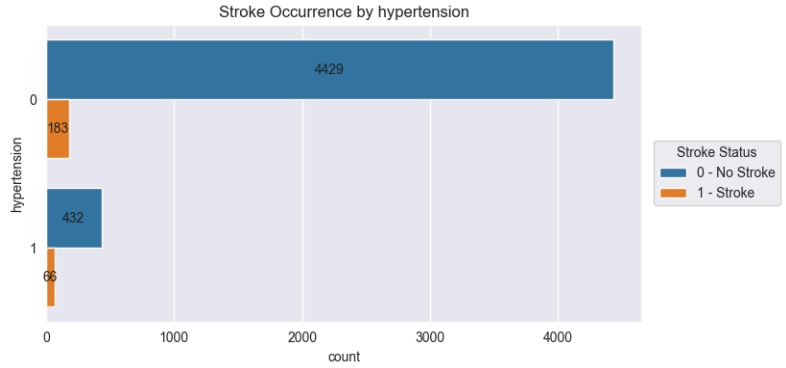
Appendix 8 Categorical Features Analysis



Stroke occurrence by gender

female 4.71%

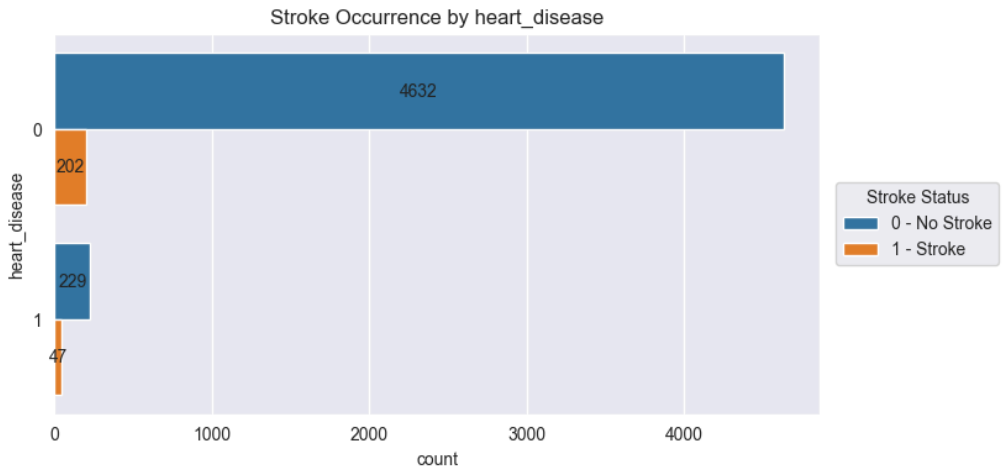
male 5.1%



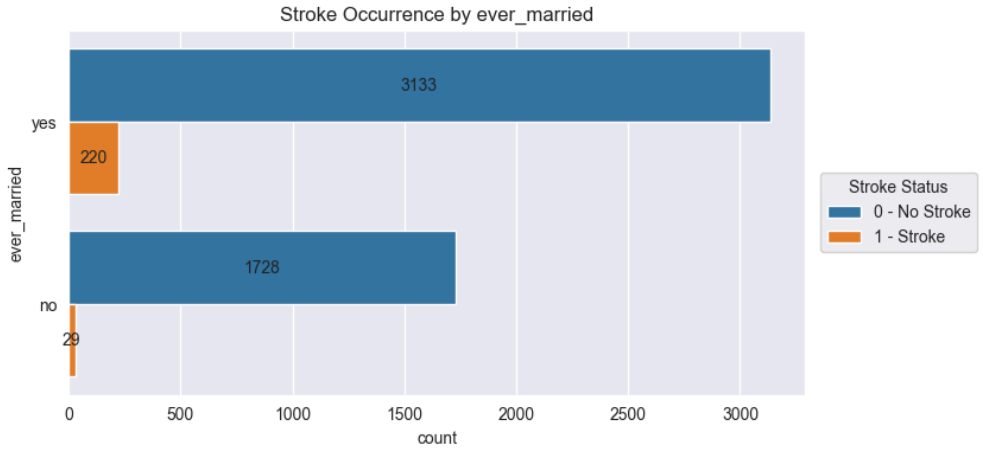
Stroke occurrence by hypertension

0 3.97%

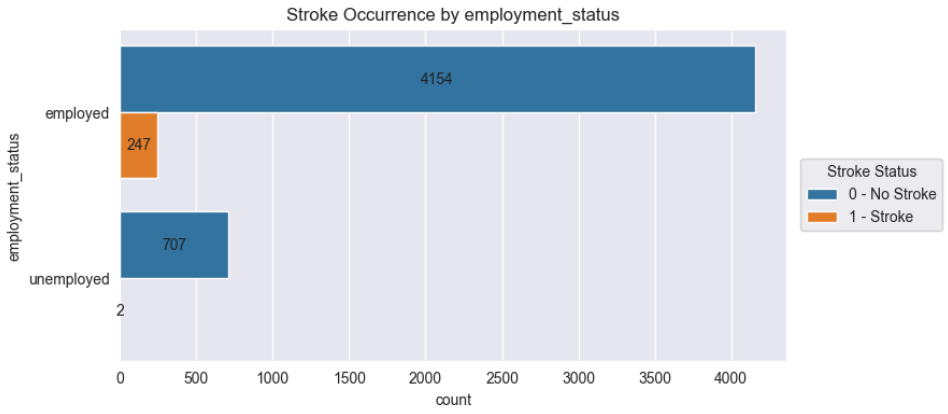
1 13.25%



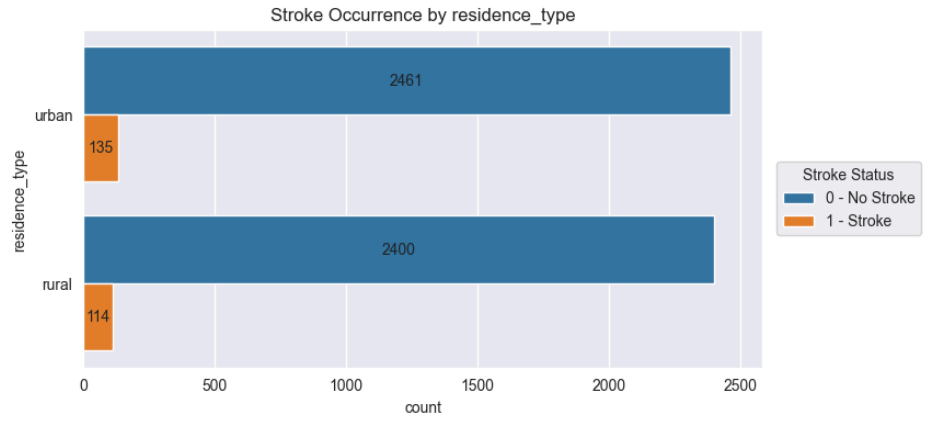
Stroke occurrence by heart_disease
0 4.18%
1 17.03%



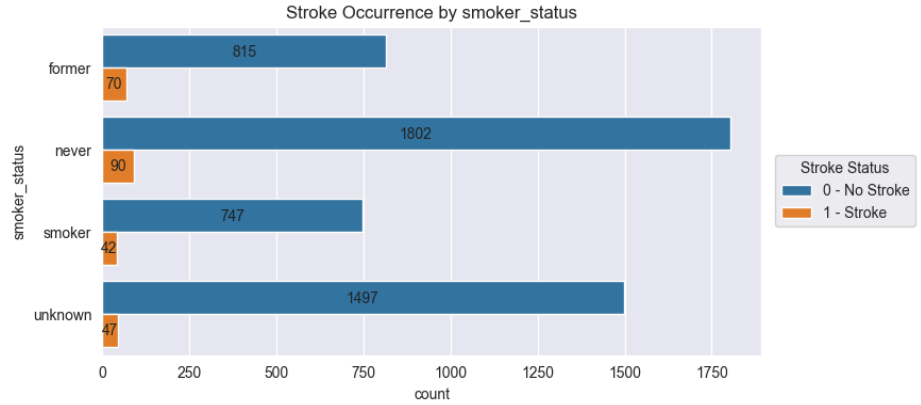
Stroke occurrence by ever_married
no 1.65%
yes 6.56%



Stroke occurrence by employment_status
employed 5.61%
unemployed 0.28%

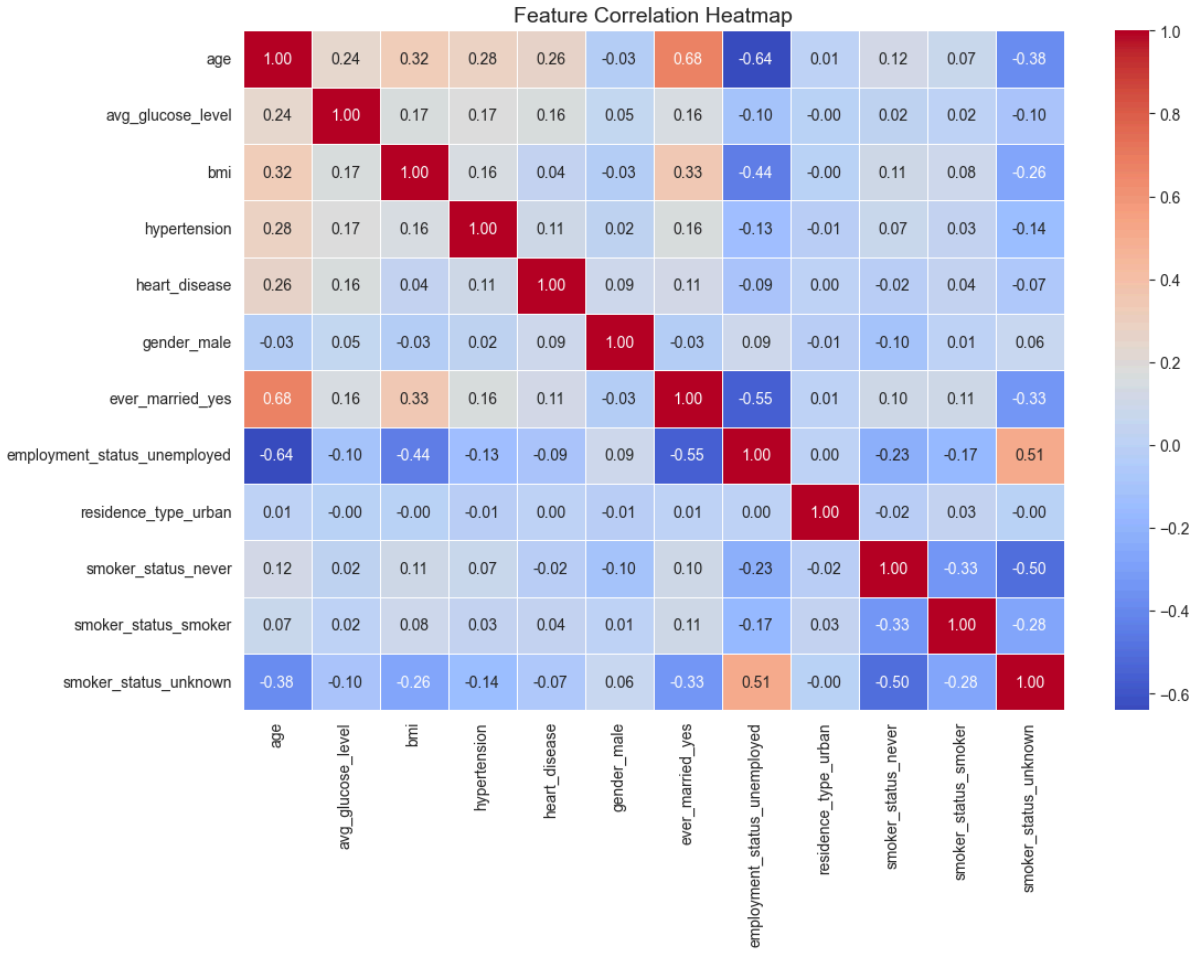


Stroke occurrence by residence_type
rural 4.53%
urban 5.2%



Stroke occurrence by smoker_status
former 7.91%
never 4.76%
smoker 5.32%
unknown 3.04%

Appendix 9: Feature Correlation Matrix



Appendix 10 Preliminary Logistic Regression Analysis

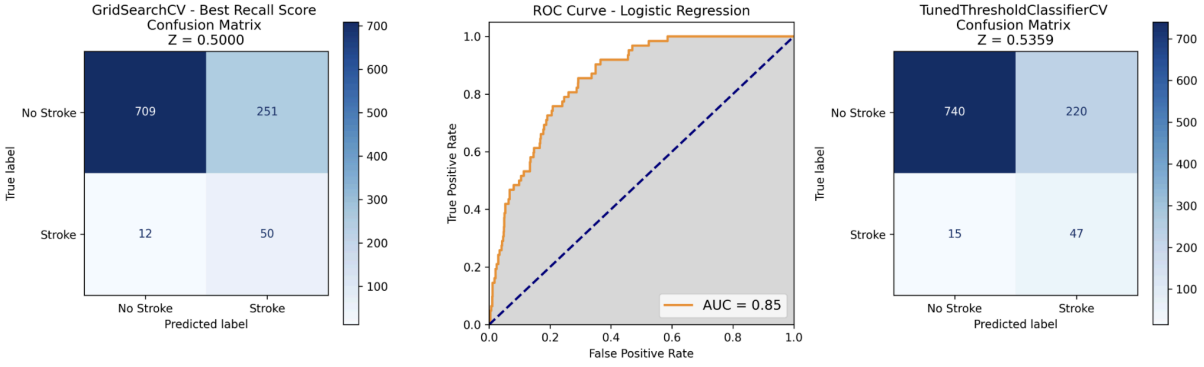
```
Optimization terminated successfully.
         Current function value: 0.155253
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                 stroke   No. Observations:                 5110
Model:                          Logit   Df Residuals:                     5097
Method:                           MLE   Df Model:                           12
Date:                Mon, 24 Mar 2025   Pseudo R-squ.:                0.2028
Time:                        16:12:13   Log-Likelihood:                -793.34
converged:                       True   LL-Null:                       -995.19
Covariance Type:            nonrobust   LLR p-value:                6.277e-79
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       -7.5523      0.587    -12.859      0.000      -8.703      -6.401
age                              0.0717      0.006     12.713      0.000       0.061       0.083
avg_glucose_level                0.0041      0.001      3.454      0.001       0.002       0.006
bmi                              0.0013      0.011      0.111      0.912      -0.021       0.023
hypertension                     0.3939      0.165      2.393      0.017       0.071       0.716
heart_disease                    0.2886      0.191      1.513      0.130      -0.085       0.663
gender_male                      0.0206      0.142      0.146      0.884      -0.257       0.298
ever_married_yes                -0.1663      0.225     -0.740      0.460      -0.607       0.274
employment_status_unemployed     0.7121      0.816      0.873      0.383      -0.887       2.311
residence_type_urban             0.0855      0.138      0.619      0.536      -0.185       0.356
smoker_status_never             -0.2017      0.176     -1.148      0.251      -0.546       0.143
smoker_status_smoker             0.1205      0.215      0.560      0.575      -0.301       0.542
smoker_status_unknown           -0.0620      0.208     -0.298      0.766      -0.470       0.346
================================================================================================
```

Appendix 11

			Model Details			Baseline Metrics			Post-Tuning Metrics		
		best_algorithm	mean_fit_time	roc_auc_score	test_recall_score	test_f1_score	test_acc_score	mod_recall_score	mod_recall_score	mod_f1_score	mod_acc_score
Approach 1: All features	Logistic Regression	LogisticRegression(C=0.01, class_weight='balan...	0.045450	0.851562	0.806452	0.275482	0.742661	0.758065	0.758065	0.285714	0.770059
	Decision Tree Classifier	DecisionTreeClassifier(class_weight='balanced'...	0.025940	0.794640	0.645161	0.271186	0.789628	0.645161	0.645161	0.299625	0.817025
	Random Forest Classifier	(DecisionTreeClassifier(max_depth=10, max_feat...	5.346738	0.847278	0.306452	0.251656	0.889432	0.854839	0.854839	0.286486	0.741683
	XGBoost Classifier	XGBClassifier(base_score=None, booster=None, c...	0.186923	0.841045	0.741935	0.288401	0.777886	0.741935	0.741935	0.284830	0.773973
Approach 2: Health-related features only	Logistic Regression	LogisticRegression(C=0.01, class_weight='balan...	0.036586	0.851798	0.806452	0.276243	0.743640	0.758065	0.758065	0.289231	0.773973
	Decision Tree Classifier	DecisionTreeClassifier(class_weight='balanced'...	0.020503	0.794640	0.645161	0.271186	0.789628	0.645161	0.645161	0.299625	0.817025
	Random Forest Classifier	(DecisionTreeClassifier(max_depth=10, max_feat...	2.905645	0.844640	0.258065	0.219178	0.888454	0.903226	0.903226	0.236287	0.645793
	XGBoost Classifier	XGBClassifier(base_score=None, booster=None, c...	0.668512	0.846925	0.725806	0.283019	0.776908	0.725806	0.725806	0.277778	0.771037

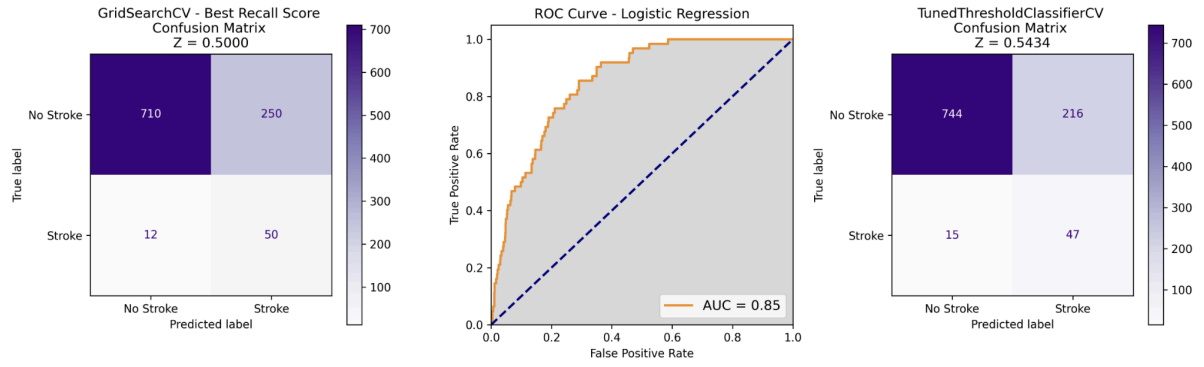
Appendix 12

Approach 1 - Logistic Regression Model: Grid Search + ROC AUC + Tuned Threshold Adjustment



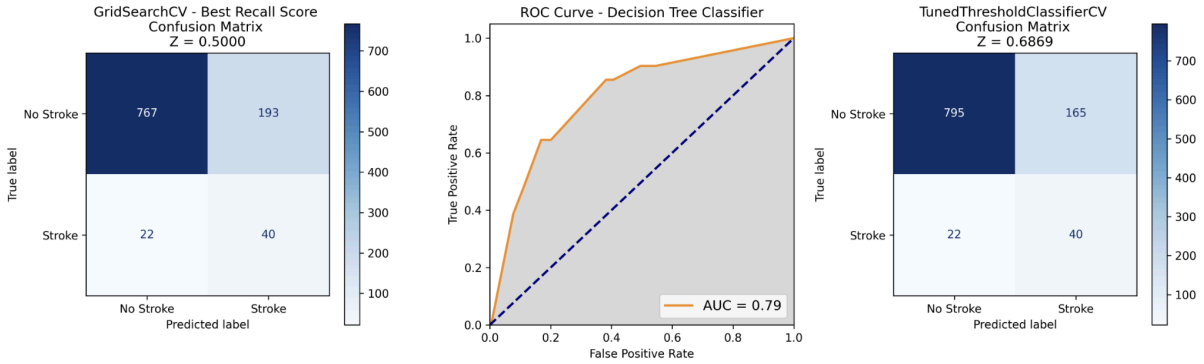
Appendix 13

Approach 2 - Logistic Regression Model: Grid Search + ROC AUC + Tuned Threshold Adjustment



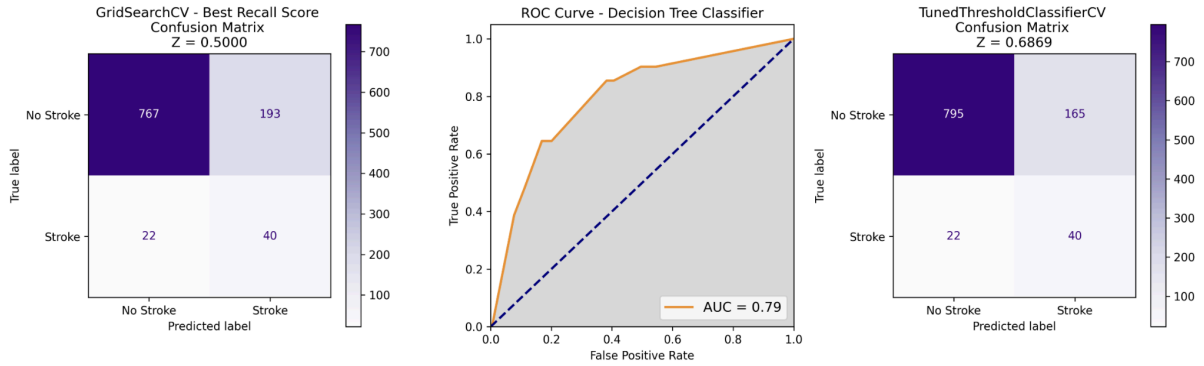
Appendix 14

Approach 1 - Decision Tree Classifier Model: Grid Search + ROC AUC + Tuned Threshold Adjustment



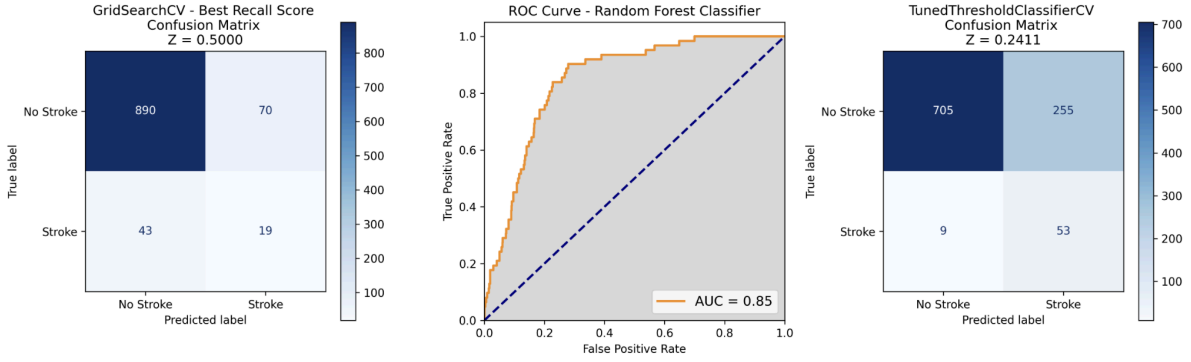
Appendix 15

Approach 2 - Decision Tree Classifier Model: Grid Search + ROC AUC + Tuned Threshold Adjustment



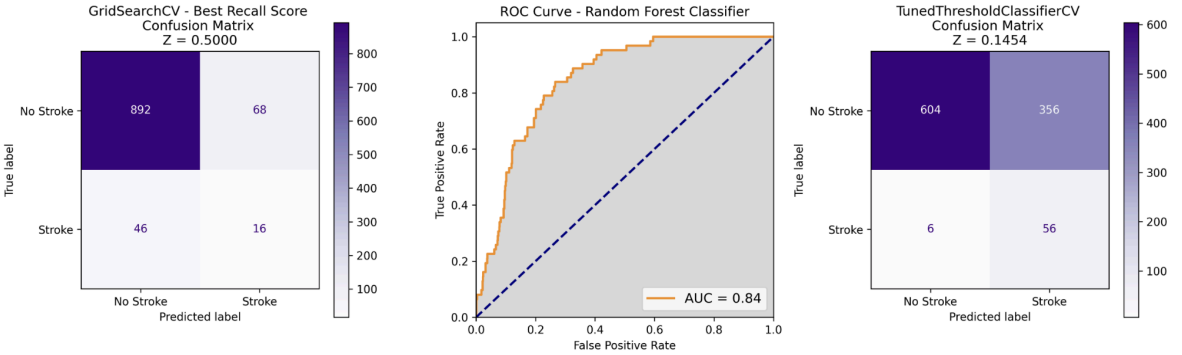
Appendix 16

Approach 1 - Random Forest Classifier Model: Grid Search + ROC AUC + Tuned Threshold Adjustment



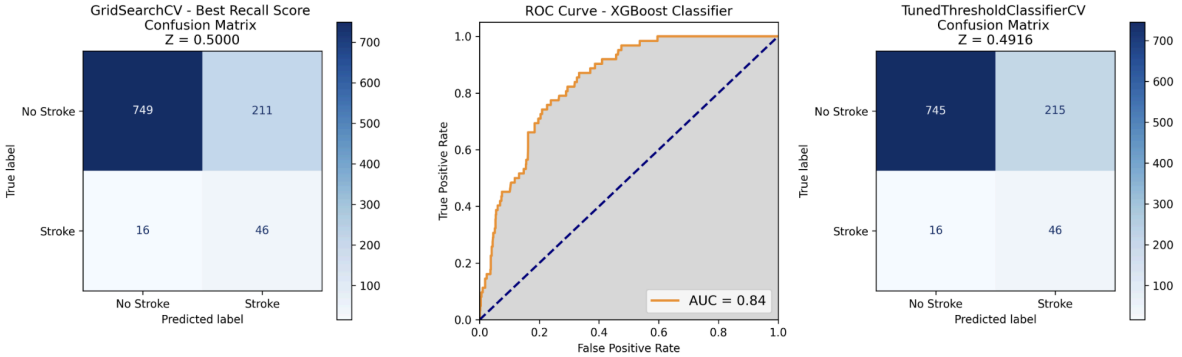
Appendix 17

Approach 2 - Random Forest Classifier Model: Grid Search + ROC AUC + Tuned Threshold Adjustment



Appendix 18

Approach 1 - XGBoost Classifier Model: Grid Search + ROC AUC + Tuned Threshold Adjustment



Appendix 19

Approach 2 - XGBoost Classifier Model: Grid Search + ROC AUC + Tuned Threshold Adjustment