Model Assessment and Explainability in Machine Learning for Stroke Risk Prediction

Coursework 1 for COMP0172: Artificial Intelligence for Biomedicine and Healthcare

Richard Huang

15 Nov 2024

Contents

Chapter 1

Introduction

Stroke is a leading global health challenge, ranking as the second-leading cause of death and the third-leading cause of combined death and disability worldwide. According to recent estimates, millions worldwide experience a stroke each year, leading to significant mortality and long-term disability. In fact, according to World Stroke Organization, (WSO)Feigin et al. [8] strokes account for one in four deaths globally, emphasizing the urgent need for improved prediction and prevention strategies. Early identification of individuals at high risk for stroke can enable timely intervention, potentially saving lives and reducing healthcare costs.

1.1 Opportunities and challenges

Recently, machine learning has emerged as a powerful tool for predictive modeling, enabling automated stroke risk assessments and assisting in early intervention strategies. By leveraging these methods, healthcare providers can achieve more accurate and efficient predictions, potentially improving patient outcomes[14]. However, several challenges persist. The quality and completeness of medical data can vary widely, impacting model performance and reliability. Additionally, class imbalance, where certain conditions (like the absence of stroke) are far more common than others, poses significant issues. High imbalance can lead to biased models that struggle to generalize effectively, ultimately reducing the model's practical value in clinical environments [10].

1.2 Related Works

In recent studies, various machine learning techniques have been explored for stroke prediction, leveraging diverse datasets and methods to improve accuracy and interpretability. Emon et al. [5] developed a weighted voting classifier using attributes from a dataset obtained from a medical clinic in Bangladesh. This model achieved a notable accuracy of 97% and an AUC score of 0.93, outperforming several state-of-the-art classifiers in comparative evaluations.

Dev et al. [4] applied Principal Component Analysis (PCA) to analyze feature importance and correlations within an Electronic Health Record (EHR) dataset. By benchmarking three popular classification models—neural networks, decision trees, and random forests—they found that the neural network performed best when paired with an optimized feature set, achieving an accuracy of 78%.

Explainable AI techniques have also been applied to enhance the interpretability of stroke prediction models. Mohammed and George [12] used methods such as SHAP and LIME to assess feature importance, providing valuable insights into model prediction dynamics. Similarly, Pamungkas, Wibawa and Cahya [13] implemented various explainable AI performance metrics to evaluate stroke prediction models, offering a more comprehensive and fair assessment of model effectiveness.

1.3 Our Approach

As discussed, class imbalance and model reliability are two critical challenges in the effective application of machine learning techniques for stroke prediction [10]. To address class imbalance, we implement the Synthetic Minority Over-sampling Technique (SMOTE) [2] [3] to augment the minority class in the target variable, thereby enhancing the model's ability to generalize across classes. A range of machine learning models is employed, including logistic regression, XGBoost, random forest, and gradient boosting. For model optimization, we use a grid search approach to identify the optimal parameter configurations for the best-performing model. To ensure a comprehensive evaluation of model performance, multiple evaluation metrics are applied. Finally, we incorporate explainable AI techniques, specifically SHAP values and Partial Dependence (PD) profiles, to analyze feature importance and enhance model interpretability across varying feature values, offering deeper insights into the model's predictive mechanisms.

Chapter 2

Methodology

2.1 Dataset

2.1.1 Data Description

The dataset consists of 5110 patient details and 12 attributes. These attributes can be majorly divided into three parts which are lifestyles factors, non-controllable factors and medical risk factors[7]. Lifestyle factors include *smoking_status* and potentially *work_type*, reflecting habits and activity levels. Medical risk factors, such as *hypertension*, *heart_disease*, and *avg_glucose_level*, represent conditions that can be managed to reduce stroke risk. Non-modifiable factors, including *age*, *gender*, and *Residence_type*, are inherent characteristics crucial for risk assessment.

2.1.2 Exploratory Data Analysis

We conduct exploratory data analysis in this section to discover some useful pattern of the dataset. From Figure 2.1, the numerical variables reveal distinct patterns and potential issues. Age shows a fairly uniform distribution, while average glucose level and BMI are right-skewed, with some high outliers.

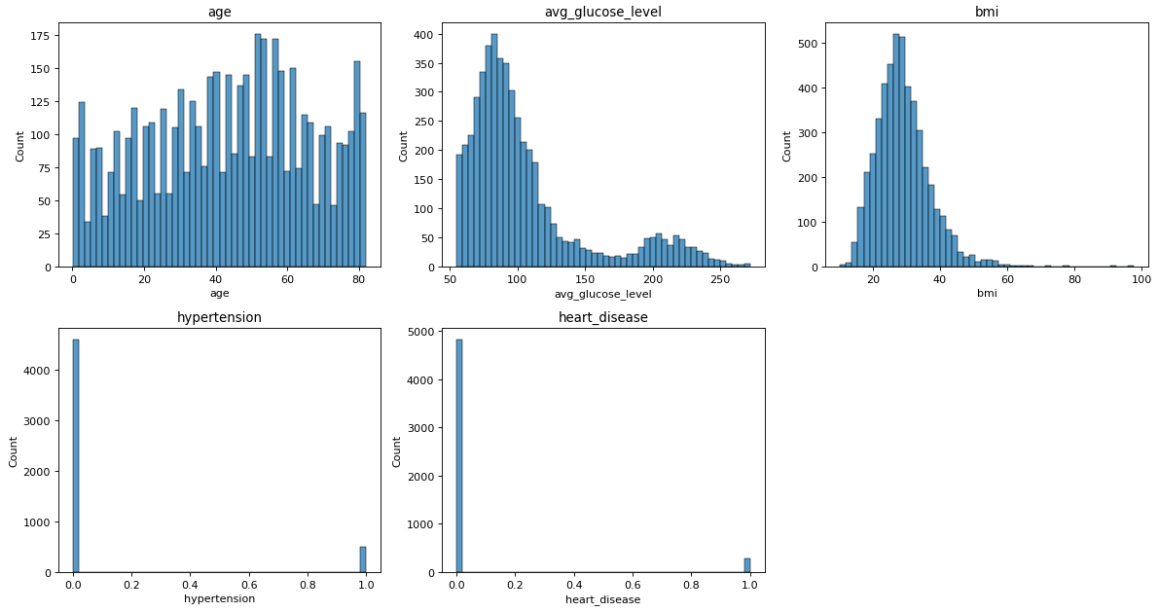

Figure 2.1: Numerical Variables Distribution

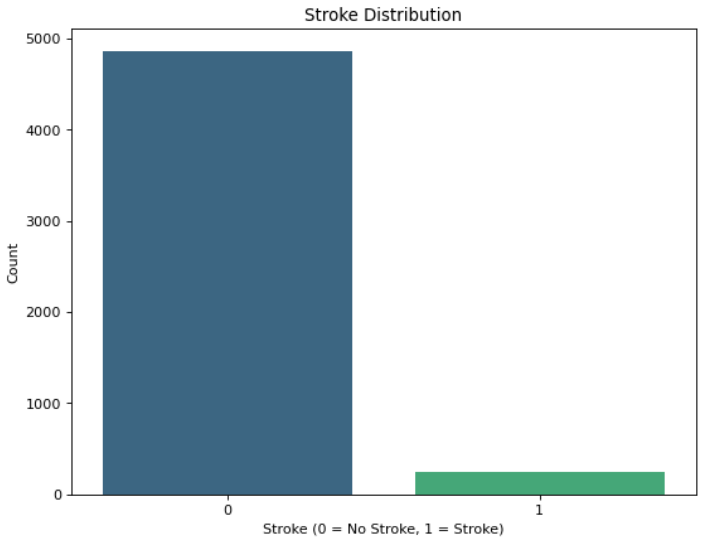

Figure 2.2: Imbalanced Target Variable

Figure 2.2 shows that the target variable (stroke) is highly imbalanced with a significant majority of cases labeled as No Stroke (0) and only a small fraction labeled as Stroke (1). This imbalance presents challenges for our model training, as it may become biased toward predicting the majority class, leading to poor performance in identifying stroke cases.

Figure 2.3: Influence of age, BMI and smoking status on storke

Figure 2.3 and Figure 2.4 show the relationships between age, BMI, and stroke occurrence by gender and smoking status. In both cases, stroke cases are more common among older individuals, suggesting age as a essential risk factor. BMI shows no clear trend with stroke across gender or smoking status, as stroke cases are scattered around the average BMI line. Smoking status also does not appear to significantly impact stroke occurrence in this data, indicating that age may be a stronger predictor of stroke than BMI or smoking habits.

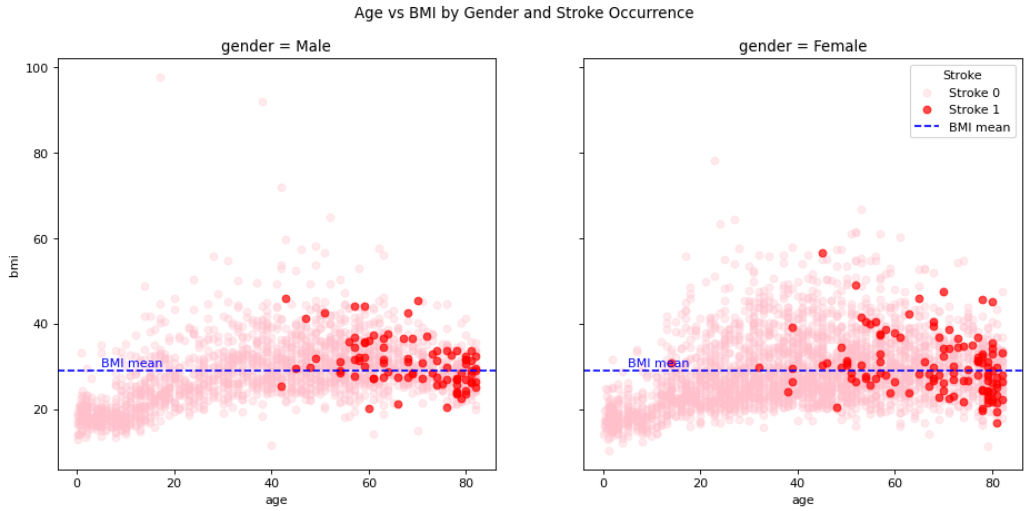

Figure 2.4: Influence of age, BMI and gender on storke

2.1.3 Data Cleaning and Preprocessing

After performing exploratory data analysis, we split the original dataset into training and testing datasets in an 80:20 ratio to avoid data leakage during preprocessing.

Data cleaning and preprocessing were conducted on the training dataset, which consisted of 4087 records. We identified the need to address outliers and handle missing values. To detect outliers, we used the interquartile range method. Abnormal values in the *avg_glucose_level* and *bmi* variables were identified as outliers and removed to ensure robust analysis.

Additionally, 3.9% of the BMI values were found to be missing. These missing values were handled using mean imputation, a straightforward method that preserves the BMI distribution without introducing significant bias.

Categorical variables in the dataset were encoded into numerical values for compatibility with machine learning models. The following transformations were applied:

- **Gender**: Male mapped to 1 and Female to 0.
- **Ever Married**: Yes mapped to 0 and No to 1.
- **Work Type:**Private, Self-employed, Govt_job, children, and Never_worked mapped to 0, 1, 2, 3, and 4, respectively.
- **Residence Type**: Urban mapped to 0 and Rural to 1.
- **Smoking Status**: formerly smoked, never smoked, smokes, and Unknown mapped to 0, 1, 2, and 3, respectively.

Lastly, the target variable demonstrates class imbalance as shown in Figure 2.2. To manage this, we used SMOTE method [2] [3] to oversample the minority class, resulting in a balanced distribution shown in Figure 2.5 for the training data.

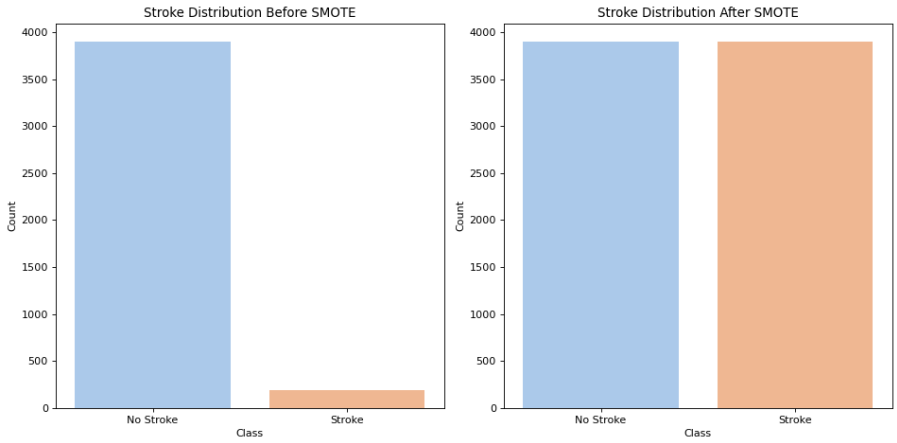

Figure 2.5: Stroke distribution before and after SMOTE oversampling

2.2 Machine Learning Models

2.2.1 Models Selection and Optimization

For this prediction task, we use four machine learning models: logistic regression, XGBoost, random forest, and gradient boosting. Logistic regression, as a linear model, serves as a baseline model due to its simplicity, efficiency, and ease of interpretability. XGBoost, an advanced ensemble learning technique, leverages gradient boosting with optimized performance and scalability, making it highly effective for handling structured data. Random forest and gradient boosting, as ensemble methods, capture complex patterns, with random forest reducing variance through averaging and gradient boosting improving accuracy iteratively.

To optimize performance, we apply grid search for hyperparameter tuning for each model to ensure the best performance. The models' performance will be evaluated and compared across different evaluation metrics introduced below.

2.2.2 Evaluation Metrics

After fitting the predictive models, we will evaluate their performance using multiple metrics:

- The F1 score, the harmonic mean of precision and recall, balances both metrics, making it useful for imbalanced datasets where minimizing both false positives and false negatives is important.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

  where

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

- Accuracy is the ratio of correctly predicted instances to the total instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total Instances}}$$

  . While it provides an overall correctness measure, accuracy may be less reliable for imbalanced data.

- AUC measures the area under the ROC curve, quantifying the model's ability to distinguish between positive and negative instances across thresholds.

- A confusion matrix summarizes model predictions as true positives, true negatives, false positives, and false negatives, providing detailed insight into prediction errors.

2.3 Explainable AI

For explainable AI technique, we apply the methods such as variable-importance measure and partial-dependence profiles from the book Burzykowski and Tomasz [1] and the open resources on GitHub from Mohammed and George [12].

Partial Dependence (PD) profiles and SHAP values are key components of Explainable AI (XAI), designed to enhance the transparency and interpretability of machine learning models. PD profiles provide global insights by showing how a model's predictions vary with changes in specific features, showcasing trends, stability, and potential boundary effects. SHAP values are a powerful tool for feature importance analysis, which quantifies each feature's contribution to model predictions. They provide global insights by ranking influential features and local explanations by attributing specific predictions to feature contributions.

Chapter 3

Result and Discussion

In the results section, we will analyze the test dataset, which was separated from the original dataset and comprises 20% of the total data, amounting to 1022 records used for evaluation purposes.

3.1 Model Performance

3.1.1 AUC

The ROC curves with AUC values for the four models are presented in Figure 3.1. Random Forest achieved the highest AUC score of 0.8484, closely followed by Logistic Regression at 0.8462. XGBoost also performed well with an AUC of 0.8387, while Gradient Boosting recorded a slightly lower AUC of 0.8270. These results indicate that Random Forest and Logistic Regression had the best performance in distinguishing between positive and negative cases, with all models exhibiting strong predictive capabilities.

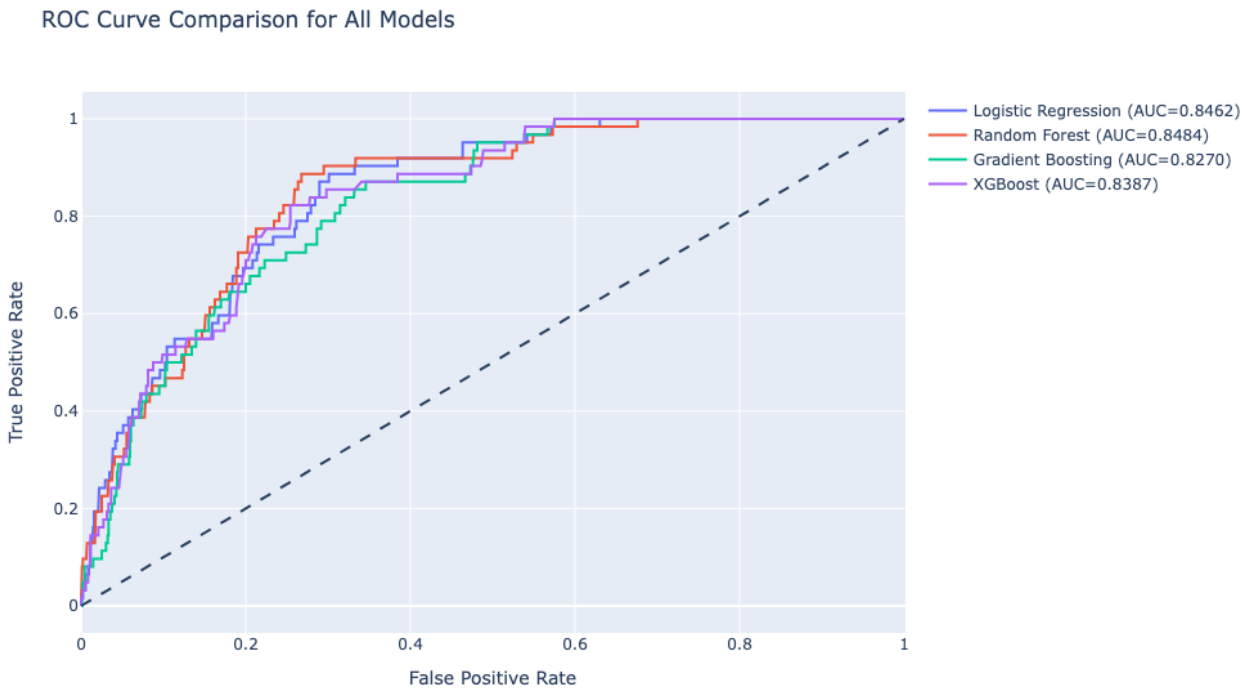

Figure 3.1: AUC for different models

3.1.2 Models Performances Statistics

The performance metrics for the four models are summarized in Table 3.1. Among these models, Random Forest achieved the highest recall (0.8871) and F1 score (0.2918), making it the most effective model for identifying true positives while balancing precision and recall. However, all models shows relatively low F1 scores, primarily due to the imbalance in the testing dataset, where the number of negative cases (non-stroke) significantly outnumbered the positive cases (stroke). Despite this, the Random Forest model still outperformed the others with the best trade-off between recall, precision, and overall classification effectiveness. Logistic Regression performed competitively with an AUC of 0.8462 but had lower recall (0.8065) and precision (0.1582) compared to Random Forest. XGBoost demonstrated good performance with an AUC of 0.8387 and relatively balanced recall (0.8226) and precision (0.1609), while Gradient Boosting recorded the lowest AUC (0.8270) and F1 score (0.2604), suggesting it is the least effective model for this task.

Model	**Recall**	**Precision**	**F1 Score**	**Accuracy**	**AUC**
Logistic Regression	0.806452	0.158228	0.264550	0.727984	0.846220
Random Forest Classifier	0.887097	0.174603	0.291777	0.738748	0.848438
Gradient Boosting Classifier	0.709677	0.159420	0.260355	0.755382	0.827008
XGB Classifier	0.822581	0.160883	0.269129	0.728963	0.838651

Table 3.1: Performance Metrics for Stroke Prediction Models

3.1.3 Confusion Matrix

The confusion matrices shown in Table 3.2 for the four models provide insights into their classification performance. Among these, Random Forest correctly identifying 55 true positives while maintaining a relatively low number of false positives (260). Its ability to minimize false negatives (7) highlights its superior sensitivity in detecting stroke cases. Logistic Regression also performed well, with 50 true positives and 266 false positives, though its higher false negative count (12) indicates slightly reduced sensitivity compared to Random Forest. XGBoost demonstrated similar performance to Logistic Regression. In contrast, Gradient Boosting had the lowest sensitivity, which identified only 44 true positives with 232 false positives, and demonstrated the highest false negative count (18).

	Actual: 0	**Actual: 1**	**Total**
Predicted: 0	694	12	706
Predicted: 1	266	50	316
Total	960	62	1022

(a) Logistic Regression

	Actual: 0	**Actual: 1**	**Total**
Predicted: 0	700	7	707
Predicted: 1	260	55	315
Total	960	62	1022

(b) Random Forest Classifier

	Actual: 0	**Actual: 1**	**Total**
Predicted: 0	728	18	746
Predicted: 1	232	44	276
Total	960	62	1022

(c) Gradient Boosting Classifier

	Actual: 0	**Actual: 1**	**Total**
Predicted: 0	694	11	705
Predicted: 1	266	51	317
Total	960	62	1022

(d) XGBoost Classifier

Table 3.2: Confusion Matrices for Stroke Prediction Models

3.2 Model Explainability

3.2.1 Feature Importance

The feature importance analysis, visualized in Figures 3.2 and Figure 3.3. Based on SHAP values in Figure 3.2, age is identified as the most significant feature, with the largest average impact on the model output, followed by average glucose level and BMI. Figure 3.3, which uses dropout loss to determine feature importance, also emphasizes *age* as the most influential variable, consistent with the SHAP-based analysis. Both methods highlight the moderate contributions of factors like marital status and smoking status, while features such as *gender* and *residencetype* show minimal impact. These findings align with similar results reported in Mohammed and George [12], which also identified age and glucose level as pivotal predictors.

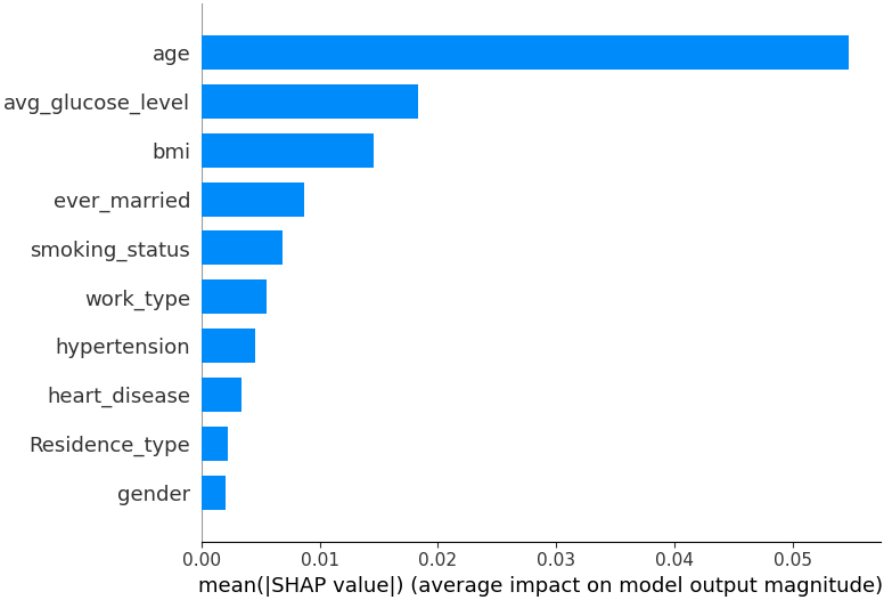

Figure 3.2: Feature Importance based on SHAP values (Code:[12])

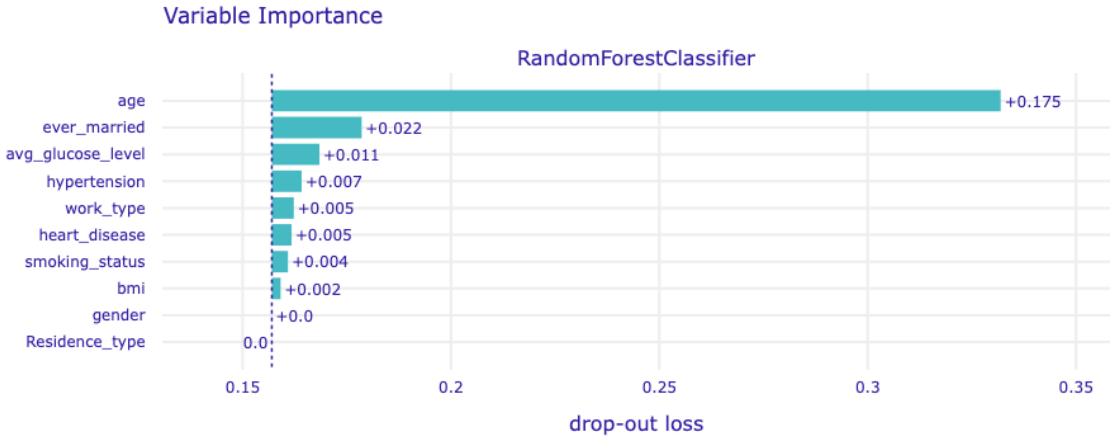

Figure 3.3: Feature Importance based on dropout loss (Code:[1])

3.2.2 PD Profiles

The PD profiles in Figure 3.4 illustrate how the four models respond to changes in *age* and *average_glucose_level*. For age, all models predict an increasing stroke risk as age rises. Logistic Regression demonstrates a smooth, linear trend, consistent with its global interpretability, while tree-based models like Random Forest, XGBoost, and Gradient Boosting capture stepwise patterns, with sharp increases around ages 40 and 60, reflecting threshold effects captured by their structures.

For *average_glucose_level*, Logistic Regression again shows a gradual rise, providing a smooth, global relationship. In contrast, tree-based models show more localized and stepwise variations. Random Forest and XGBoost display relatively stable predictions across most glucose levels.

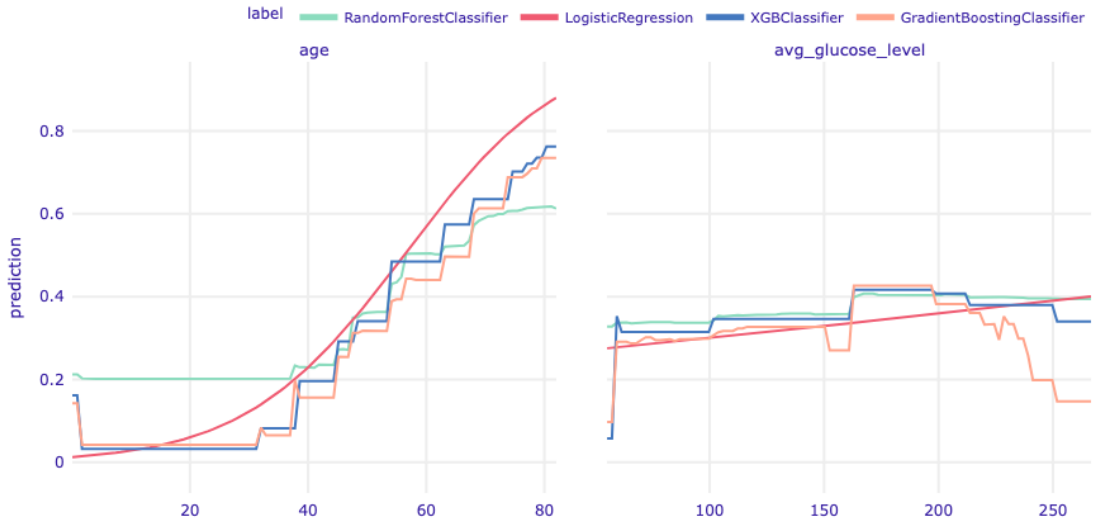

Figure 3.4: PD profiles for numerical data (Code:[1])

The PD profiles in Figure 3.5 show how categorical variables influence stroke risk in the Random Forest model. Being married (Yes), having hypertension, or heart disease is associated with slightly higher predicted stroke risk. Moreover, the plot indicates that categorical variables have a relatively small impact on stroke risk predictions in the Random Forest model, as the changes in prediction values along the y-axis are minimal, typically ranging between 0.32 and 0.38. This suggests that while variables like hypertension, marital status, and smoking status contribute to predictions, their overall influence is limited compared to key numerical features like *age* and *avg_glucose_level*.

Figure 3.5: PD profiles for categorical data (Code:[1])

Chapter 4

Conclusion and Discussion

4.1 Summary

This report develops and evaluates machine learning models for stroke risk prediction, addressing imbalanced datasets with SMOTE and optimizing models like Random Forest, Logistic Regression, Gradient Boosting, and XGBoost. Random Forest demonstrated the best performance with the highest AUC and F1 score. Explainable AI techniques, including SHAP values and Partial Dependence profiles, were applied to interpret feature importance, indicating age and glucose level as key predictors. Despite these strengths, the model occasionally fails by misclassifying stroke cases, particularly false negatives, which could have critical implications. There is room for improvement by incorporating additional features, refining feature engineering, and continuously monitoring model performance to adapt to real-world clinical scenarios.

4.2 Potential Challenges

In light of these findings, it is crucial to consider the challenges of deploying stroke prediction models in clinical settings. Deploying machine learning models in healthcare environments presents significant obstacles, particularly regarding data quality, bias, and integration. Clinical datasets are often fragmented and institution-specific, necessitating extensive preprocessing to ensure models perform reliably across diverse populations [9] [15]. Moreover, algorithmic bias inherited from historical data poses a risk of unfair outcomes for underrepresented groups, showcasing the need for robust validation. The black box nature of many models further complicates adoption, as it can undermine trust among clinicians and patients, which essentially require explainable AI approaches to enhance transparency and confidence [6] [11]. Additionally, dynamic clinical environments introduce model drift, requiring continuous monitoring and updates to maintain consistent performance [9].

Workflow integration and regulatory hurdles present further barriers. Embedding models seamlessly into clinical workflows demands user-friendly interfaces and compatibility with electronic health records, without disrupting clinical routines[11]. Ethical considerations, such as maintaining patient privacy and ensuring compliance with safety and liability standards, are also critical[15]. Addressing these challenges requires robust infrastructure, ongoing evaluation, and clear regulations to ensure that machine learning models, like those for stroke risk prediction, can be effectively deployed to improve healthcare delivery[6].

Bibliography

[1] Przemyslaw Biecek Burzykowski and Tomasz. *Explanatory Model Analysis*. URL: https://ema.drwhy.ai/.

[2] N. V. Chawla et al. 'SMOTE: Synthetic Minority Over-sampling Technique'. In: *Journal of Artificial Intelligence Research* 16.16 (June 2002), pp. 321–357. DOI: https://doi.org/10.1613/jair.953.

[3] Damien Dablain, Bartosz Krawczyk and Nitesh V. Chawla. 'DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data'. In: *IEEE Transactions on Neural Networks and Learning Systems* 34.9 (Sept. 2023), pp. 6390–6404. DOI: https://doi.org/10.1109/tnnls.2021.3136503.

[4] Soumyabrata Dev et al. 'A predictive analytics approach for stroke prediction using machine learning and neural networks'. In: *Healthcare Analytics* 2 (Nov. 2022), p. 100032. DOI: https://doi.org/10.1016/j.health.2022.100032.

[5] Minhaz Uddin Emon et al. 'Performance Analysis of Machine Learning Approaches in Stroke Prediction'. In: *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (Nov. 2020), pp. 1464–1469. DOI: https://doi.org/10.1109/iceca49313.2020.9297525.

[6] Pouyan Esmaeilzadeh. 'Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations'. In: *Artificial Intelligence in Medicine* 151 (May 2024), p. 102861. DOI: https://doi.org/10.1016/j.artmed.2024.102861.

[7] Günter Fahrnberger. *Distributed computing and internet technology : 15th international conference, ICDCIT 2019, Bhubaneswar, India, January 10-13, 2019 : proceedings*. Cham Springer, 2019, pp. 238–252. ISBN: 9783030053659.

[8] Valery L Feigin et al. 'World Stroke Organization (WSO): Global Stroke Fact Sheet 2022'. In: *International Journal of Stroke* 17.1 (Jan. 2022), pp. 18–29. DOI: https://doi.org/10.1177/17474930211065917.

[9] Steve Harris et al. 'Clinical deployment environments: Five pillars of translational machine learning for health'. In: *Frontiers in Digital Health* 4 (Aug. 2022). DOI: https://doi.org/10.3389/fdgth.2022.939292.

[10] Tianyu Liu, Wenhui Fan and Cheng Wu. 'A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset'. In: *Artificial Intelligence in Medicine* 101 (Nov. 2019), p. 101723. DOI: https://doi.org/10.1016/j.artmed.2019.101723.

[11] Charles Lu et al. *Deploying clinical machine learning? Consider the following... **. 2023.

[12] K. Mohammed and G. George. 'IDENTIFICATION AND MITIGATION OF BIAS USING EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) FOR BRAIN STROKE PREDICTION'. In: *Open Journal of Physical Science (ISSN: 2734-2123)* 4.1 (Apr. 2023), pp. 19–33. DOI: https://doi.org/10.52417/ojps.v4i1.457.

[13] Yuri Pamungkas, Adhi Dharma Wibawa and Meiliana Dwi Cahya. 'Electronic Medical Record Data Analysis and Prediction of Stroke Disease Using Explainable Artificial Intelligence (XAI)'. In: *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control* (Nov. 2022). DOI: https://doi.org/10.22219/kinetik.v7i4.1535.

[14] Aaron N. Richter and Taghi M. Khoshgoftaar. 'A review of statistical and machine learning methods for modeling cancer risk using structured clinical data'. In: *Artificial Intelligence in Medicine* 90 (Aug. 2018), pp. 1–14. DOI: https://doi.org/10.1016/j.artmed.2018.06.002.

[15] Angela Zhang et al. 'Shifting machine learning for healthcare from development to deployment and from models to data'. In: *Nature Biomedical Engineering* 6.12 (July 2022), pp. 1330–1345. DOI: https://doi.org/10.1038/s41551-022-00898-y.