SUBJECT NAME: Predictive Analytics in Business
SUBJECT CODE: DBA3803
ACADEMIC YEAR: AY 2023/2024, SEMESTER 1

Name	Matriculation Number
Lin WenYan	A0241093M
Lim Ker Yin	A0241030E
Van Erven Oliver Daniel	A0278460X
Darren Darius Tan Jia Wei	A0173269W

Sectional Group: A1 (Group 4)

Section 1: Introduction to Dataset and Objective

Dataset Overview

The stroke prediction dataset includes diverse input parameters related to individual patients, encompassing demographic details (such as gender and age), medical history (indications of different diseases), and lifestyle factors (like smoking status). Each record in the dataset represents an individual patient and their associated health and demographic details.

Objective

The primary objective is to develop a predictive model that can accurately identify individuals with an elevated risk of experiencing a stroke. This model will leverage the available dataset to discern patterns or characteristics associated with an increased likelihood of stroke. We aim to optimise the False Negative Rate (FNR) in our model, balancing the need for accurate identification of high-risk individuals with the risk of overfitting the model, which will be demonstrated in Section 6.

Significance of Project

Stroke is a major global health issue, being the second leading cause of death worldwide as per the World Health Organization (WHO). It accounts for approximately 11% of total deaths globally (WHO, 2020). Early prediction of stroke can significantly impact patient outcomes, potentially saving lives and reducing long-term disabilities (Chen et al., 2023). Identifying individuals at high risk of stroke enables timely medical interventions and lifestyle modifications.

Section 2: Overcoming Imbalanced Data and Resultant Samples

In the original dataset, comprising 5110 instances, only 249, or less than 5%, are positive cases of stroke. This imbalance poses a challenge for predictive modelling, as models trained on such data might be biased towards the majority class and perform poorly in identifying the minority class, which in this case is the critical class of interest. A stratified shuffle split is employed to maintain the same stroke class distribution in both train and test datasets.

Prior to applying the Synthetic Minority Over-sampling Technique (SMOTE), the class distribution in the training set was heavily skewed, with only 199 instances of the minority class (label '1' - Stroke) compared to 3889 instances of the majority class (label '0' - No Stroke). After applying SMOTE, the class distribution became perfectly balanced, with each class constituting exactly 50% of the dataset. This balance is vital for training machine learning models, as it helps to avoid bias towards the majority class and improves the model's ability to predict the minority class accurately.

The subsequent tests of models, both with and without SMOTE, revealed improved performance when SMOTE was applied. This aligns with expectations as SMOTE enhances model accuracy by addressing class imbalance.

Section 3: Initial Model Structure

The initial approach to modelling involves the construction of a simple and interpretable model with a focus on the following aspects shown in the table below.

Data	Continuous features: Age, Glucose level, Body Mass Index (BMI)
	Discrete features (Number of categories): Gender (2), Marital status (2), Hypertension (2), Heart disease (2), Work type (5), Residence (2), Smoking Status (4)
	Target Variable (Class label): Stroke (1) or No stroke (0)
Loss	Gini impurity
Structure	Classification tree
Constraint	Maximum depth

Based on the simple classification tree model in Appendix 1, the depth range is adjusted to an optimal maximum depth of 4 to enhance interpretability. The decision tree suggests that age is the primary factor in predicting stroke, with subsequent decisions based on employment type and smoking status, among other factors. While the model shows a good fit to the training data as evidenced by the Area Under the ROC Curve (AUC) score of 75.91%, its performance on the test data suggests it is less effective, especially concerning the F1 score of 18.12%, which is critical for medical prediction models. It would be crucial to look into improving the model by feature engineering and trying different models to improve the prediction outcome.

Section 4: Changing of Data and Constraints with Ensemble Methods

The analysis involved training various models, including Logistic Regression, Decision Trees, Bagged Decision Trees, Random Forest, Support Vector Machine (SVM), Extra Trees, and Gradient Boosting. Each model was initialised with a consistent random state to ensure reproducibility and fairness in comparison. The Gradient-Boosting model demonstrated superior performance, achieving an AUC of approximately 80.78%. This elevated AUC score suggests that Gradient Boosting is particularly effective in differentiating between stroke and non-stroke cases, a critical capability in medical predictive modelling. Given its superior performance in this key metric, Gradient Boosting has been selected for more in-depth analysis in this study.

Section 5: Candidate Splitting with Gradient Boost

In the development of the gradient boost model, measures were taken to control the number of candidate splits (or “cuts) by consolidating categories within features through domain knowledge pertaining to this dataset. For instance, the distinction between 'formerly smoked' and 'currently smokes' may not be as significant as the distinction between having smoked at all versus never having smoked. Consequently, 'Formerly Smoked' and 'Smokes' were combined into a single category. This aggregation reduces the complexity of the model by decreasing the number of categories the algorithm needs to consider when making splits at each node, which can help prevent overfitting and improve computational efficiency. For clarity, the ‘Children’ and ‘Never Worked’ categories were merged to distinguish individuals who were unemployed. This systematic approach to feature engineering not only simplifies

the model but also incorporates domain knowledge to potentially enhance the model's predictive accuracy and interpretability. Factor encoding was deemed essential for the model, as the interpretability is important to determine the most significant feature that may cause stroke. As shown in Appendix 2, it is difficult to interpret the model, particularly in determining how each category within the labels affects the prediction outcome.

Through encoding, the SHapley Additive exPlanations (SHAP) values depicted in Appendix 3 clearly explain the contribution of each feature to the predictive model's output. The plot indicates that 'Age' is the most significant predictor, with its higher values typically contributing to an increased model output, which could suggest a greater likelihood of the event the model predicts. 'Smoking status', both current and past, along with 'Residence type' and 'Gender', show a substantial influence on the model's predictions, indicated by their spread across the SHAP value axis. The colours, likely representing feature value magnitudes, allow us to see that higher (red) and lower (blue) values of these features have varying effects on the predictions. However, features typically associated with high medical risk, such as heart disease and hypertension, exhibit lower SHAP values. This counterintuitive result could be due to multicollinearity, where these conditions are correlated with other variables (like age or BMI) that have already been accounted for by the model. As a result, the individual impact of heart disease and hypertension is diminished in the presence of these correlated features.

Section 6: Refining to the Best Loss Function that corresponds to Domain Knowledge

This section focuses on identifying the best loss function that effectively predicts stroke cases, with a particular emphasis on lowering the False Negative Rate (FNR). Although dummy encoding enhances model interpretability, the Catboost model demonstrates a generally lower FNR. Given the objective of minimising the FNR, the Catboost model is selected for refining the loss function.

Baseline Model

The baseline model, trained without class weighting, served as a reference point for comparison with other techniques. It exhibited a high accuracy of approximately 94% and an AUC score of 0.81 but struggled with a high FNR of about 94%. This indicates a significant shortfall in detecting actual stroke cases, a serious concern in medical diagnostics, despite having a low False Positive Rate (FPR) of 1.23%.

Inverse Frequency Weighting

To address the imbalance, the next step involves using inverse frequency weighting, assigning more weight to the minority class (stroke cases). As compared to the baseline model, this method decreased the model's accuracy to 78% and maintained a similar AUC of 0.80. Importantly, it considerably improved the FNR to about 34%, substantially reducing the number of missed stroke cases. However, this improvement came with an increased FPR of 21.19%, indicating a higher rate of false alarms.

Grid Search

Grid search is employed to systematically explore different combinations of class weights including ratios 1/10, 1/20/, 1/30, and 1/40. The best performance is achieved with a weight ratio of 1:10 (non-stroke to stroke), yielding the highest AUC of 0.993 among the tested ratios. Making use of the best weight ratio of 1:10 for prediction does not improve the AUC scores from the baseline model, however, improved the FNR slightly to 64% and also raised the FPR to 11%.

Optimising Threshold Based on Domain Knowledge

While the primary goal of the prediction model is to minimise the FNR, setting the threshold to zero, thereby predicting a stroke for every case, is not a viable solution. Such an approach would lead to significant overfitting of the model. Instead, thresholds should be thoughtfully

established, reflecting the clinical context and the consequences of decisions based on risk stratification (Wynants et al., 2019).

Wynants et al. (2019) suggest that it is more appropriate to assign cost values to different outcomes like False Negative (FN), False Positive (FP), True Positive (TP), and True Negative (TN) and use these values to determine an optimal risk threshold. This method is particularly crucial in healthcare settings, where the costs of misdiagnosis, whether as an FN or an FP, are substantial in terms of both patient health and economic impact. In the context of stroke prediction, for example, the costs associated with an FN (a missed stroke diagnosis) are significantly higher than those of an FP (unnecessary medical procedures due to overdiagnosis), reflecting the grave health consequences of failing to detect a stroke. We Incorporated the costs into our model with justifications shown in the table below, setting a threshold helps to balance the risk of FNs and FPs in a way that is most beneficial for patient outcomes and healthcare efficiency.

Outcome	Stroke Occurs (True Condition)	No Stroke (True Condition)
Predicted Stroke (Model Prediction)	20 (C_TP): Costs reflect necessary treatment and rehab but are minimised by the benefits of early intervention.	10 (C_FP): Cost accounts for the economic and psychological impacts of unnecessary medical procedures and healthcare inefficiencies due to overdiagnosis (Lafata et al., 2004)
Predicted No Stroke (Model Prediction)	95 (C_FN): High costs associated with acute stroke management and rehabilitation (Lucas-Noll et al., 2023), combined with the potential for severe health consequences, including disability and death (Johns Hopkins Medicine, n.d.).	No cost as the prediction is accurate and no intervention is needed.

With the threshold adjustment by directly applying the given formula, the model's performance changes significantly as compared to the baseline model. The new model has a lower accuracy of about 71% and an AUC score of about 0.76. However, the FNR is substantially reduced to approximately 36%, indicating fewer missed stroke cases. This comes at the cost of a higher FPR of about 29%.

Optimising Threshold Based on Custom Loss Function with Domain Knowledge

A custom loss function with domain knowledge was used to experiment with finding the best threshold, minimising test loss. The optimal threshold was set at 0.408 which minimises the custom loss. This resulted in an optimal model with an accuracy of about 86% and an AUC of approximately 0.76. Similar to optimise threshold with domain knowledge, optimise threshold based on custom loss function helps to reduce FNR to 60% but increases the FPR to 11.63% as compared to the baseline model.

In real-world scenarios, especially in medical contexts, the trade-off between an FNR and an FPR is critical. Reducing the FNR is crucial because missing a stroke case can have severe or fatal consequences. On the other hand, a high FPR can lead to unnecessary medical interventions and anxiety for patients. Both the Inverse Frequency Weighting and Optimising Threshold Based on Domain Knowledge methods significantly improved the FNR to 34% and 36%, respectively. However, Inverse Frequency Weighting also achieved a marginally lower

FPR of 21.10% compared to 20%, making it more suitable for medical diagnostics. This method prioritises reducing the FNR, aligning with the critical need to identify as many true stroke cases as possible. Furthermore, from a cost perspective, the implications of a false negative in stroke prediction are significantly higher than those of a false positive, encompassing not just financial costs but also the impact on patient health, quality of life, and healthcare resources. Hence, this strategic emphasis on minimising FNR in stroke prediction underscores the importance of tailoring model optimisation strategies to the specific needs and consequences in the field of application, especially in high-stakes areas like medical diagnostics.

Section 7: Experimental Changes to Linear Model Structure

To increase the clarity and understandability of the predictive model, logistic regression was chosen as a more interpretable alternative to the gradient boosting method. Initially, a basic logistic regression model was trained, yielding the following metrics: an accuracy of approximately 94.91%, an AUC score of 0.85, an FPR of about 0.21%, and an FNR of 100%. The 100% FNR clearly pointed out that the model was consistently predicting only the '0' class, indicating no stroke occurrence. This pattern of prediction highlighted a serious case of model underfitting, where the model failed to capture the complexity and nuances necessary to identify actual stroke cases.

In an attempt to refine the model's predictions, a Lasso penalty was incorporated into the logistic regression. The expectation was that this regularisation technique would help in managing overfitting and improving model generalisation. However, this adjustment did not yield the desired improvement. Although there was a slight uptick in accuracy to about 95.11% and an improved AUC score of 0.86, the FNR remained stubbornly at 100%. This persistence of the FNR underscored a continuing challenge in model underperformance, where the model was still unable to identify any true positive stroke cases.

In response to these challenges, the next step involved a more nuanced approach – the implementation of a custom loss function with an adjusted threshold within the logistic regression framework. This adjustment resulted in an accuracy of 82.38%, an AUC of 0.86, an FPR of 17.08%, and a FNR of 28%. Notably, this approach significantly reduced the FNR, indicating a substantial enhancement in the model’s ability to correctly identify positive cases. This outcome was particularly significant because it not only demonstrated the logistic regression model’s improved reliability but also showcased its ability to outperform the more complex gradient boost model with Inverse Frequency Weighting. The success of this adjusted logistic regression model lies in its effective balance of the trade-off between FPs and FNs, a crucial factor in medical diagnostics, particularly in predicting conditions as critical as strokes.

The SHAP summary plot for the logistic regression model, as detailed in Appendix 4, highlights 'Age' as the most influential feature, similar to our previous findings above. A concentration of pink dots to the far right illustrates that higher ages are strongly associated with an increased likelihood of the model predicting a stroke. In stark contrast, 'work_type_children' and 'Residence_type_Rural' tend to skew to the left, indicating that these features are likely to decrease the model's prediction of a stroke occurring. The distribution of SHAP values for 'heart disease' and 'hypertension' is notably less dispersed and situated lower on the graph. Again, this phenomenon was further explained in Section 5. The summary plot also reveals the nuanced interplay between features. For example, 'smoking_status_never smoked' and 'Residence_type_Urban' show a mix of SHAP values on both sides of the zero line. This indicates that the presence or absence of these features can either positively or negatively influence the model's output, depending on how they combine with other variables in the data.

Section 8: Governance of Overall Model with Hyperparameter Tuning

In the advanced stage of model refinement, hyperparameter tuning plays a crucial role in optimising the performance of both the gradient boosting and the Lasso logistic regression models. The primary goal here is to enhance model accuracy and reliability while minimising the risk of overfitting. To achieve this, a strategic approach is employed: employing grid search within a cross-validation framework.

With hyperparameter tuning on the logistic regression, it leads to an accuracy of 77%, an AUC score of 0.86, an FPR of 23% and an FNR of 16%. Notably, there is a marked improvement in the FNR compared to the initial Lasso logistic regression model with a custom loss function. This enhancement indicates a significant reduction in the number of stroke cases that the model fails to detect, thereby increasing its clinical utility. Such improvements are a testament to the efficacy of hyperparameter tuning in refining the model's predictive capabilities. By meticulously adjusting the model parameters, the risk of misclassifying stroke cases as non-cases (FNs) is substantially mitigated, which is particularly crucial in medical diagnostics where the cost of missed diagnoses is high. The advanced hyperparameter-tuned model emerges as a robust tool, potentially offering more reliable and actionable insights for healthcare practitioners.

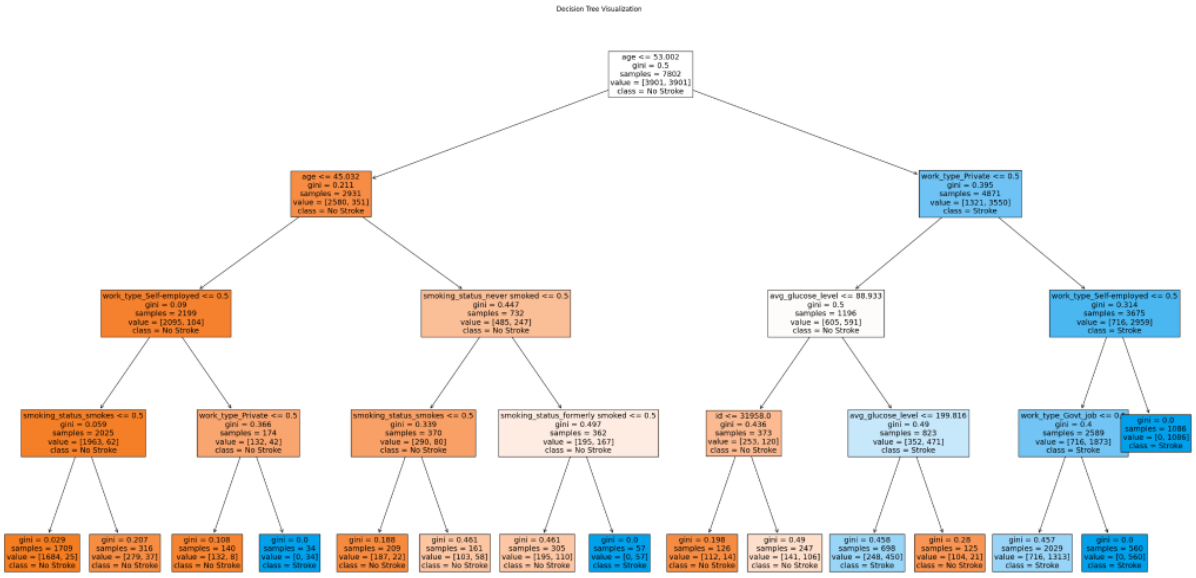
In the quest to refine the predictive capabilities of the gradient boost model, hyperparameter tuning was identified as a critical step in model enhancement. Due to computational time issues, a validation set was added instead of cross-validation, which decreased the overall validation set size for the tuning of the gradient boost model. However, a discrepancy in performance metrics between the validation and test sets highlighted a case of overfitting: while the model achieved an impressive AUC of 0.99 on the validation set, this figure significantly decreased to 0.78 on the test set, and the FNR increased dramatically from 13% to 81%. To address this, regularisation parameters were intensified, which helps to penalise complexity and encourage simpler models that may generalise better. Alongside this, the learning rate was decreased, and the depth of the trees was reduced, both of which are strategies aimed at curbing overfitting. Despite these efforts, the AUC remained at 0.99 for the validation set and decreased further to 0.77 for the test set, with the FNR remaining high at 81% for the test set, suggesting that overfitting was still a concern. Future improvements could involve enlarging the validation set, which may offer a more robust assessment of model performance and a better approximation of generalisation error. Additionally, a more nuanced and focused effort on hyperparameter fine-tuning could lead to more substantial improvements. Such fine-tuning often requires a delicate balance and a deep understanding of the model's behaviour. Currently, the enhancement of the gradient boost model through grid search is proving to be time-consuming. This limits the ability to experiment with and iteratively refine the model's hyperparameters extensively.

Section 9: Conclusion

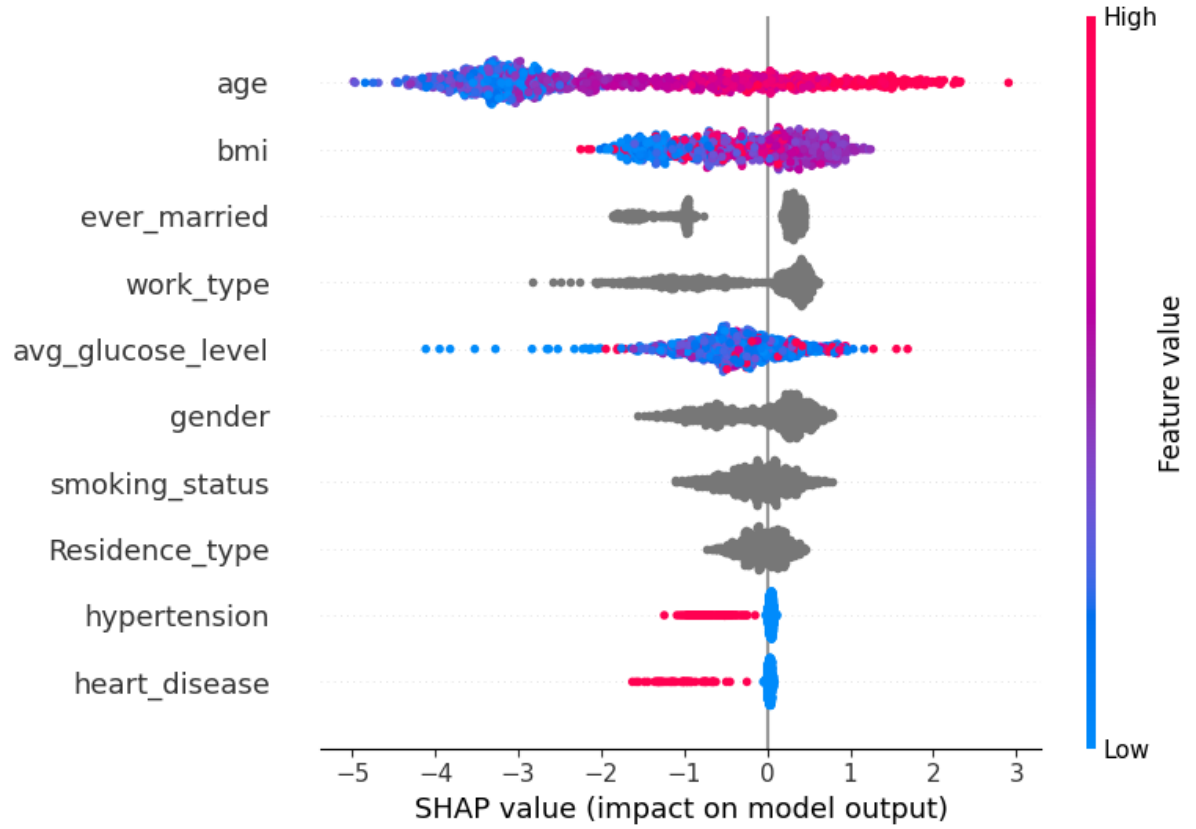
This study embarked on an extensive journey to develop a predictive model capable of accurately identifying individuals at high risk of stroke. Through meticulous data analysis, balancing of imbalanced datasets, and comprehensive model testing, we have managed to refine our predictive capabilities significantly. The application of techniques such as SMOTE, inverse frequency weighting, and custom loss functions, alongside strategic hyperparameter tuning, has yielded promising results. By modifying the threshold and weights based on our domain-specific loss considerations, we have successfully reduced the incidence of FNs relative to the models described in Sections 5 to 7. After refining the gradient boost and logistic regression models, they have demonstrated substantial improvement in their predictive accuracy. However, with more computational capacity, we have the potential to further refine the gradient boost and logistic regression models through more extensive hyperparameter tuning, promising even greater improvements in their predictive accuracy.

Appendix

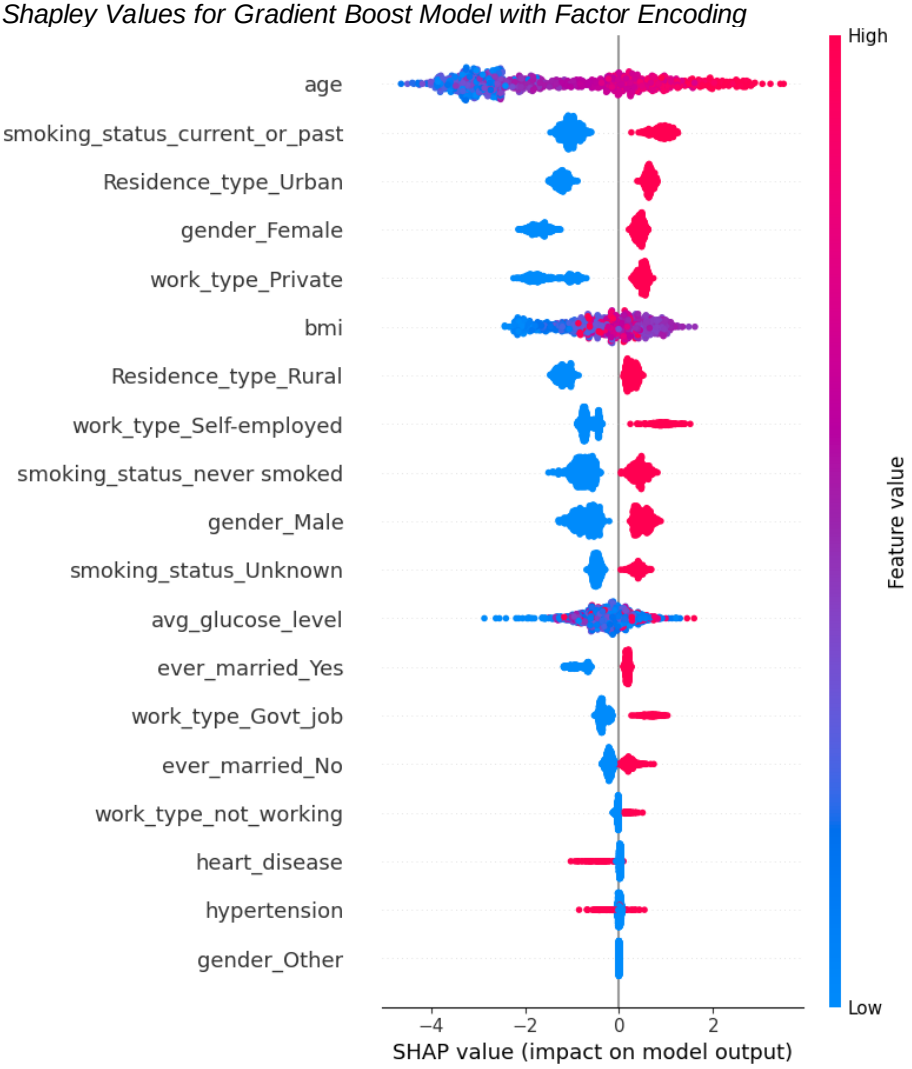
Appendix 1: Classification Tree



Appendix 2: Shapley Values for Gradient Boost Model with No Factor Encoding



Appendix 3: Shapley Values for Gradient Boost Model with Factor Encoding



Appendix 4: Shapley Values for Logistic Regression

References

Chen, Y.-W., Lin, K., Li, Y., & Lin, C.-J. (2023, February 23). Predicting patient-reported outcome of activities of daily living in stroke rehabilitation: a machine learning study. *Journal of NeuroEngineering and Rehabilitation*, *20*, 25. https://doi.org/10.1186/s12984-023-01151-6

Johns Hopkins Medicine. (n.d.). *Effects of Stroke*. Johns Hopkins Medicine. Retrieved November 22, 2023, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/effects-of-stroke

Lafata, J. E., Simpkins, J., Lamerato, L., Poisson, L., Divine, G., & Johnson, C. C. (2004, December). The economic impact of false-positive cancer screens. *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *13*(12), 2126-2132. https://pubmed.ncbi.nlm.nih.gov/15598770/

World Health Organization (WHO). (2020, December 9). *The top 10 causes of death*. World Health Organization (WHO). Retrieved November 21, 2023, from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Wynants, L., Smeden, M. v., McLernon, D. J., Timmerman, D., Steyerberg, E. W., & Calster, B. v. (2019, October 25). Three myths about risk thresholds for prediction models. *BMC Med*, *17*, 192. https://doi.org/10.1186/s12916-019-1425-3