Patient Stroke Prediction

CIS 575 Project Report

Table of Contents

Table of Figures

Executive Summary

Medicine is not an exact science and often requires physicians to use their own knowledge to predict the type of treatment best suited for their disease. Beyond that, two people with exact physiological characteristics (age, weight, cholesterol, etc.) could present with completely different diseases or lack thereof. In this report the goal will be to use a variety of patient details (age, weight, medical histories, etc.) to build complex models that can help predict if a patient will suffer from a stroke. An advantage of this type of analysis is that doctors usually have the patient information needed for these models, so it could allow the physician to make objective assessments of their patients' risk for a future stroke. This will allow for more targeted treatment for higher-risk patients and less unnecessary treatment for lower-risk patients. There is a particular interest in stratifying the patient risk estimated in our models. Treating all patients who have an increased risk with the most aggressive treatment would not always be in the best interests for all patients due to the potential of unwanted side effects of certain treatments. However, under treatment of those who have increased risk carries the danger of them suffering from a potentially preventable stroke. Being able to stratify patient risk would help to eliminate these two extremes.

Business Problem

According to the CDC strokes are a leading cause of death for adult Americans, and even if the patient survives, it can lead to lifelong and debilitating disability. The best way to help reduce deaths by stroke is to begin a course of treatments with anticoagulants or antiplatelets as early as possible. Given an easy-to-use calculation which delivers patient stroke risk, physicians can better estimate the type of risk their patients have to a stroke and provide treatment with a more targeted approach.

Business Objective

There currently exist a tool used by physicians called the CHADS-VAS Score[1] which calculates a patient risk score between zero and nine. Using this score a healthcare provider (H.P) will have a better idea of how aggressive a treatment their patient needs. This tool uses age, biological gender, presence of congestive heart failure, hypertension, history of strokes, heart disease and

[1] 1: Used for patients with atrial fibrillation.

diabetes to predict. Creating a model that uses similar variables in addition to adding social determinants of health such as marriage status, work type, and residence type may provide a better prediction than is currently being used.

Data Source and Adjustments

The source of the data was taken from the website Kaggle.com. Specifically, this dataset was originally uploaded by a user named “fedesoriano” and is described as real but confidential. Not a lot of information is known about the source, but this dataset was created 26-Jan-2021 and has been used numerous times for similar analysis. The user that posted this dataset is named Federico Soriano and is currently a Data Scientist in Madrid, Spain.

In the dataset there were 10 independent variables, and the dependent variable was encoded as a 0 or 1. Before beginning any exploration the ID column was dropped from the dataset since it held no bearing in the subsequent analysis. The row IDs created in R were used as a form of index. The dataset was made up of 3 numerical variables and 7 categorical variables.

Figure 1: Variables in Dataset

Dependent Variable	Definition
Stroke	0 = No Stroke; 1 = Stroke

Independent Variables		Definition
ID	(Numerical)	Patient ID
Gender	(Categorical)	Male or Female
Age	(Numerical)	Age of Patient
Hypertension	(Categorical)	0 = No Hypertension; 1 = Hypertension
Heart Disease	(Categorical)	0 = No Heart Disease; 1 = Heart Disease
Married	(Categorical)	No or Yes
Work Type	(Categorical)	Children, Govt_job, No Work, Private, Self-Employed
Residence Type	(Categorical)	Rural or Urban
Average Glucose	(Numerical)	Average glucose level in blood
BMI	(Numerical)	Body Mass Index
Smoking Status	(Categorical)	Formerly Smoked, Never Smoked, Smokes, or Unknown

The following are notable specifications about the dataset that became important during the analysis and subsequent model creation:

- There’s a total 5,110 and of those 856 were of patients under the age of 18

- Of the 856 patients under the age of 18, only 2 have had a stroke.
- BMI has a total of 201 missing values in the original dataset
- One patient has a gender of “Other”
- Work_type has 21 “never_worked” cases of which 16 are of patients under the age of 18
- Smoking Status has 1,544 cases of “Unknown” of which 682 are for patients under the age of 18.
- The original dataset is highly imbalanced, with only 4.87% of cases having a stroke

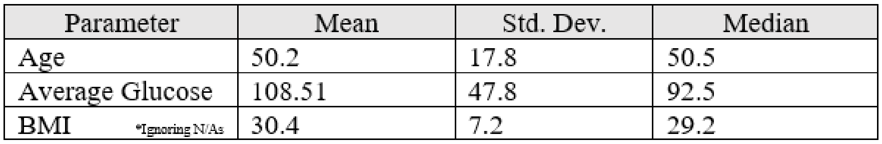
The decision was made to filter out any patient that was under the age of 18. This was primarily driven because the risk factors associated with stroke in adolescents and children, such as genetic bleeding disorders are not captured by this dataset. This decision had the added benefit of helping ease the imbalance of our response variable (5.81% vs 4.87%) as well as dropping many of the unknown or null observations. The new dataset, referred to as “adult_data” consists of 4,254 patients and has 181 missing BMI values, 5 work_types marked as “never_worked”, and 862 cases of “Unknown” smoking_status. For ease of analysis, all categorical variables were turned into factors within R.

Preliminary Data Exploration

The first step in analyzing the data is seeing how our data is distributed and what statistics they hold. Below is a summary of the adult_data dataset:

Figure 2: Variable Statistics

Parameter	Proportions
Stroke	Stroke: 5.8% - No Stroke: 94.2%
Gender	Female: 60.6% - Male: 39.4% - Other: ~0%
Hypertension	Hypertension: 11.7% - No Hypertension: 88.3%
Heart Disease	Heart Disease: 6.5% - No Heart Disease: 93.5%
Married	Not Married: 21.2% - Married: 78.8%
Work Type	Never_worked: 0.12% - Govt_job: 15.3% Private: 65.6% - Self Employed: 18.9%
Residence Type	Rural: 49% - Urban 51%
Smoking Status	Formerly Smoked: 20.2% - Never Smoked: 41.2% Smoke: 18.3% - Unknown 20.2%



Parameter	Mean	Std. Dev.	Median
Age	50.2	17.8	50.5
Average Glucose	108.51	47.8	92.5
BMI *Ignoring N/As	30.4	7.2	29.2

These statistics will be important to keep in mind when analyzing each variable's relationship with strokes. Looking closer at each parameter, age and BMI seem to be fairly symmetrical; both having means and medians close to each other. BMI is highly skewed with a skew measure of 1.23, indicating a right-skewed distribution. Looking at the categorical parameters, both heart disease and hypertension are highly unbalanced but the proportion of stroke within each level may give some insight into the impact of those conditions.

Figure 4: Hypertension and Stroke

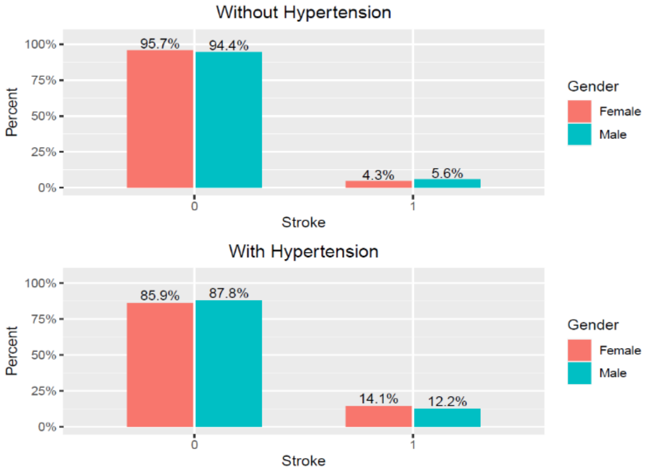
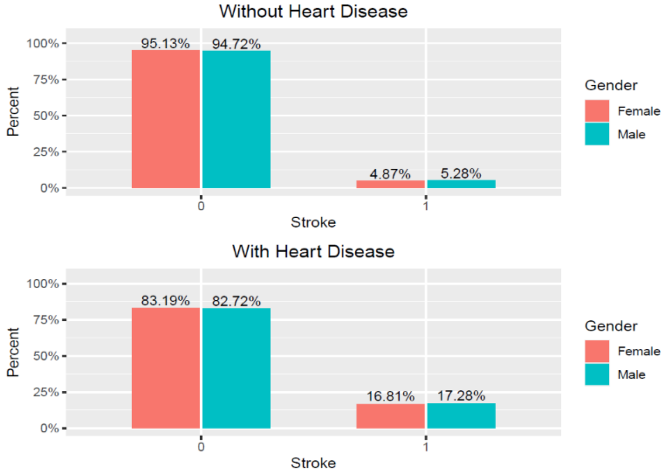


Figure 3: Heart Disease and Stroke



The figures above show 4 different groups and the prevalence to which they are vulnerable to a stroke. Since there wasn't an even split between males and females, they were analyzed separately to see how they may differ in different scenarios. In the case of with or without hypertension, females had an odds ratio of 3.65 and males had an odds ratio of 2.34. In both males and females, there was a strong association between stroke and having hypertension. Looking at heart disease, females had an odds ratio of 3.95 and males had an odds ratio of 3.75. In both cases there is a strong association between stroke and hypertension or heart disease. Females have a higher odds ratio in both cases which may point to strokes being more prevalent in women.

Box plots were also used to determine whether age, average glucose level (AGL), and BMI were strong determinates of strokes. Starting with age in **Figure 14** of the appendix, there seemed to be a strong relationship between age and strokes. While possible to live a long life without a stroke, the majority of the stroke patients are over the age of 60, with the mean patient age being 68 years old. In contrast, of the cases with no stroke, the mean age was 49. Within **Figure 14** there is also a log scaled AGL. It was scaled because of the skew the distribution presented (1.46 vs 0.845). Although there may be some relationship between the AGL and stroke, it doesn't present as a strong one in our boxplot. Finally, in **Figure 14**, BMI and stroke are plotted. This is the weakest relationship among the 3 variables in **Figure 14** and it's expected that at least in this dataset, BMI will not be a useful covariate for the models. Of all three boxplots there also doesn't seem to be a difference in how each covariate is impacted by gender, pointing to the fact that there may not be any interaction terms between gender and age, AGL, and BMI.

The last set of variables that will be explored are the categorical parameters Smoking Status, Residence Type, Ever Married, and Work Type. This was done using the proportions of each category that have a stroke. For example, looking at the Ever Married, of all the cases that have never been married, 3% have had a stroke. Of those that have been married, 6.6% have had a stroke. This means in the variable Ever Married their stroke proportions break up into 31% Never Married and 69% Married. It's important to also look at the proportions of each category level to paint a whole picture. Looking at **Figure 15** Stroke per Married, the proportion of Not Married with Stroke is 31% and the total Not Married observations represent 21.2% of the cases. This is indicative of a positive correlation between having never been married and having a stroke. There appears to be a small indication an Urban residence is correlated to Stroke as well. Looking at Smoking Status, Never Smoked represents 41.2% of the cases but only has a proportion equaling 21.5% of the total proportions. This represents a strong indication that patients who have never smoked are less likely to suffer from a stroke.

It's important to repeat the old adage "Correlation is not causation". It's improbable that getting married makes someone healthier directly, but more likely it is a sign of confounding variables. Potentially, someone who is married lives a healthier lifestyle or the combined income allows for better access to health insurance, leading to better long term health outcomes. When looking at social determinants of health to improve predictability, the main goal is to average out the types of behaviors and environments seen by these variables. Having heart disease will have a direct impact on the chance of blood clot formation thus resulting in an ischemic stroke, whereas living in an Urban area may on average lead to less healthy lifestyle. Assumptions that may be incorrect about one patient could arise, but the overall goal of these models is to optimize the treatment outcomes of at-risk patients.

Data Transformation

Now that there’s a clearer picture of how each of our variables impacts stroke risk, it is pertinent to transform our data before introducing it to our models. Our current data filtered out any patient under the age of 18, leaving 4,254 total cases. The first goal is to simplify the categorical variables with extremely rare levels. The first assumption was that biological sex is what is important medically when assessing risk and thus the 1 “Other” would need to be labeled “Female” or “Male”. This case was transformed to “Female” since it is the level most commonly present in the data set. Following that, “Never_worked” was also transformed to “Private” as it represents 2/3rd of the entire dataset. We are accepting the risk these are mislabeled because they represent 0.12% of the total cases and work type did not show to have an overwhelming impact on stroke risk during data exploration.

At this point the categorical variables have been simplified to get rid of rare levels and the next goal is to impute the missing BMI values. Ideally the imputation is done on the training data and the same transformation is done on the testing data to prevent data leakage. Unfortunately, the “trainControl” function in the R package caret was not working properly and would error when imputation was added to the pre-processing variable. The decision was made to impute the entire dataset using the “mice” function from the R package “mice”, a multivariate imputation package. This was considered low risk because as seen in the exploratory section, BMI did not have an obvious impact on Stroke, and the missing values represented 4.3% of the total cases. A RandomForest classifier was used to impute the missing BMI values because of its ability to produce unbiased estimates and is believed to perform well on datasets involving incomplete patient data [1]. This function will impute the data “m” times using gibbs sampling, in our case 5, and create 5 complete data sets for the columns with the missing data. The user can then inspect the different imputation distributions and compare them to the original data. **Figure 16** in the appendix demonstrates this for BMI. If satisfied with the imputation distributions, the mean of the 5 simulation is taken and used as the imputed result for that missing value.

The complete data was then taken, and all categorical variables were turned to dummy variables. A level was dropped as a column to keep the data from being over specified. Additionally, to avoid errors that were encountered during model creation, the Stroke factors were recoded to be “No_Stroke” and “Stroke”. It appears the train function was creating variable names based on the factor level during the preprocessing and R cannot take numbers as variable names. At this point the data was split into a training set that will be used for cross validation and test set that will be used in the end to give an honest performance review of our models. The split was 85/15 and was stratified on Stroke because of the imbalance. This resulted in a training set with 3,616 total cases and 210 positive cases and a test set with 541 cases and 31 positive cases. The choice was made to skip standardizing or scaling the data because none of the models tested are impacted by the difference in scales, only the convergence speed of our logistic model. Additionally, it would make the inferences taken from the logistic model harder to understand unless converted back.

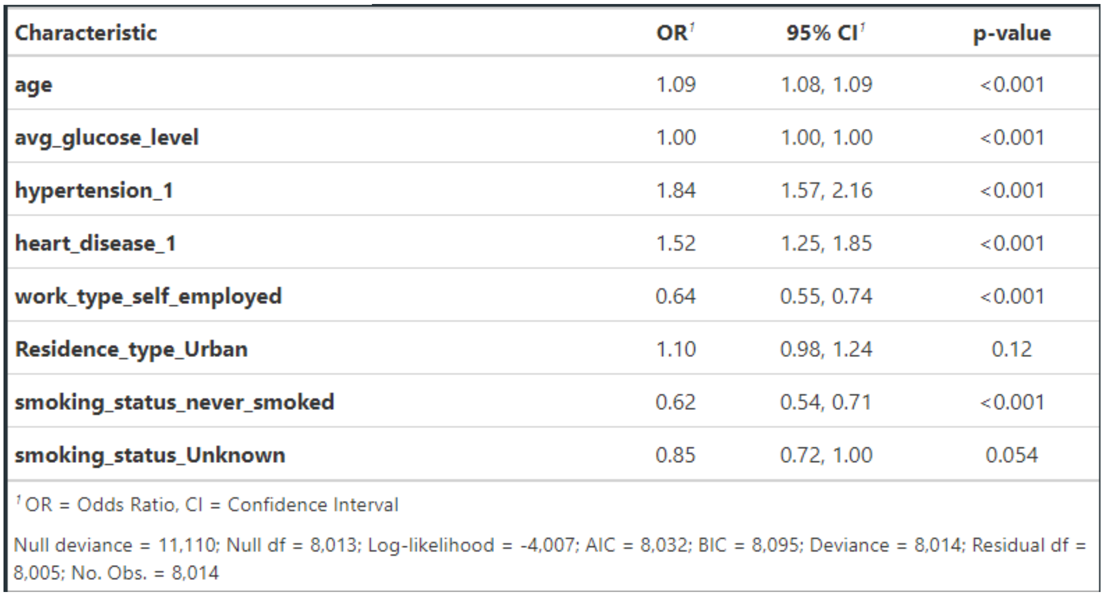
The last thing that had to be taken care of was the imbalance of the response variable. SMOTE was chosen as the means to balance the data set and particularly it was run exclusively on the training data each fold during cross validation. This was done to avoid data leakage (i.e test data having information about the training data). A KNN algorithm was used which defaults to n = 5 and it created a 50/50 response balance. In the end during our model analysis, the predicted probabilities were adjusted to match the true sample proportion of the response rate.

Model Creation and Analysis

Logistic Regression Model

The chosen model for the logistic regression was specifically a binomial logit model with an assumed dispersion coefficient of 1. Since this model doesn’t have a built-in way of dropping parameters a preliminary model was created to determine what parameters to use during the cross validation (CV) 10-fold split. The entire data set was balanced using SMOTE temporarily and the model was trained on this balanced data. It’s important to note this was not saved to the training data that was used for the CV of this and the other models. The function “stepAIC” was run against the GLM which was fit using all the covariates. This returned a summary of the statistically significant covariates that were found as seen in **Figure 17**. The first problem with this list of covariates is that partial categorical levels were dropped (e.g formerly smoked, and smokes are missing even though only 1 should be dropped by dummy columns). To solve this problem, the two categorical variables that dropped partial levels were grouped together. For work type “Private” and “Govt_job” was combined to be “GP”. This was now a binary variable that represented whether the patient worked for someone else (GP) or for themselves (Self_Employed). Additionally, smoking status was transformed to “Smoked”, “never_smoked”, and “Unknown”. Once there was a better idea of what parameters the logistic model would find significant, a 10-fold CV with SMOTE sampling was performed. The method used was the “glmStepAIC” which will do a backward model selection and select the variables with the optimal AIC. The CV uses the Cohen Kappa metric to select the best fold to use for the final model. **Figure 5** shows how the model coefficients break down, and at first glance it seems hypertension has the strongest measure of association among the categorical values. On the other end, never smoking is the most impactful category in terms of reducing risk of stroke. It also appears that Urban residence type is statistically insignificant but still optimized the AIC measure. The biggest advantage of using Logistic Models is the inference that can be pulled from the model and particularly the probabilities. In **Figure 6** below, the partial dependence plots of the 2 numerical variables and surface plots for the most impactful variables are shown. This allows for a deeper understanding of how these parameters impact stroke probability. Note these probabilities have been adjusted to the true response probabilities. At a quick glance it is apparent age exponentially increases probability of a stroke. Keeping all things equal, an 80-year-old will have a probability 22% higher than a 20-year-old to suffer from a stroke (approximately 120x higher risk). Looking at how hypertension impacts stroke risk when coupled with age, it’s clear there is significant increase in risk if the patient has hypertension. As

Figure 5: Logistic Model Summary

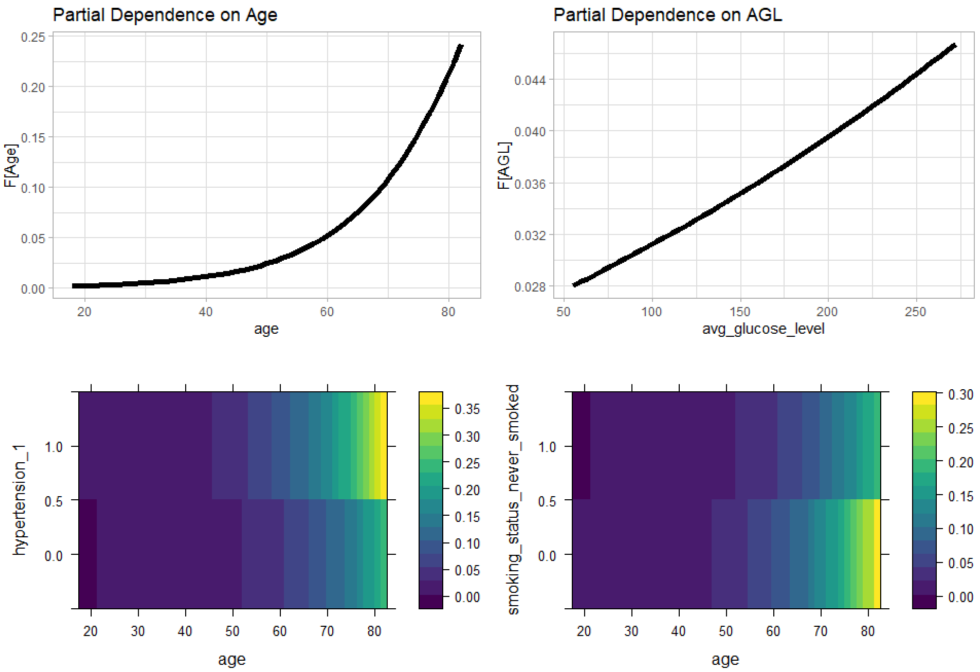


Characteristic	**OR**[1]	**95% CI**[1]	**p-value**
age	1.09	1.08, 1.09	<0.001
avg_glucose_level	1.00	1.00, 1.00	<0.001
hypertension_1	1.84	1.57, 2.16	<0.001
heart_disease_1	1.52	1.25, 1.85	<0.001
work_type_self_employed	0.64	0.55, 0.74	<0.001
Residence_type_Urban	1.10	0.98, 1.24	0.12
smoking_status_never_smoked	0.62	0.54, 0.71	<0.001
smoking_status_Unknown	0.85	0.72, 1.00	0.054

[1] OR = Odds Ratio, CI = Confidence Interval

Null deviance = 11,110; Null df = 8,013; Log-likelihood = -4,007; AIC = 8,032; BIC = 8,095; Deviance = 8,014; Residual df = 8,005; No. Obs. = 8,014 an example, comparing a 65-year-old with and without hypertension, the patient with hypertension will have a 71.2% increased risk of suffering from a stroke (7.3% vs 12.5%).

Figure 6: Logistic Model – Partial Dependence Plots



The probability predicted during CV (adjusted for true response rate) can be seen below broken up into 9 quantiles. This will be referenced in the conclusions section in order to stratify the patients to determine appropriate treatment.

Quantile	10%	20%	30%	40%	50%	60%	70%	80%	90%
Probability	0.28%	0.51%	0.86%	1.42%	2.28%	3.67%	5.93%	10.47%	18.55%

Random Forest Classifier

Continuing with the strategy established in the logistic model, the data set with the simplified categories was used. The largest difference in the Random Forest (RF) model creation was the hyperparameter tuning. A 10-fold CV strategy with smote sampling was executed. This time two models were created, the first one was created to optimize the “ntree” which dictates how many trees to grow and “mtry” which is how many predictors are sampled for splitting at each node**[1]**. The results for each of the 5 splits can be seen in **Figure 18**. The optimized values of ntree = 1000 and mtry = 1 were used to train our final RF model with a 10-fold CV. The best fold was chosen using the Cohen Kappa metric. A useful analysis can be done by plotting the accuracy

1: In R code, RF model 1 only has mtry being optimized and ntree is set to 1000. This is because the initial optimization took over 15 minutes so the results for ntree were taken and used to optimize mtry to show how the grid selection was set up. Ntress = c(500, 1000, 1500, 2000) was used. Additionally, only 5 folds are used in model 1, and 10 were used on the final parameter selection.

decrease vs gini decrease as a means of understanding the variable importance for our RF model. The further to the top right a point is, the more impactful the variable was for this model. **Figure 7** shows that age was by far the most important variable. On the other end was BMI in the bottom left corner. This is consistent with the data exploration findings as well as with the logistic model variables. It's important to note RF models can't be used as a means of inferring probability but dependence plots can still be used to gauge impact of each variable with the risk of stroke. Similar to the logistic dependance plots, risk grows much quicker as age increases, pointing to a non-linear relationship. Another major difference in the creation of this model is the use of parallel computing using the "allowParrallel" in the caret trainControl package. This added setting was able to cut the training time of the first model in half and was used in the third type of model as well, Gradient Boosting.

Figure 7: RandomForest Importance Plot

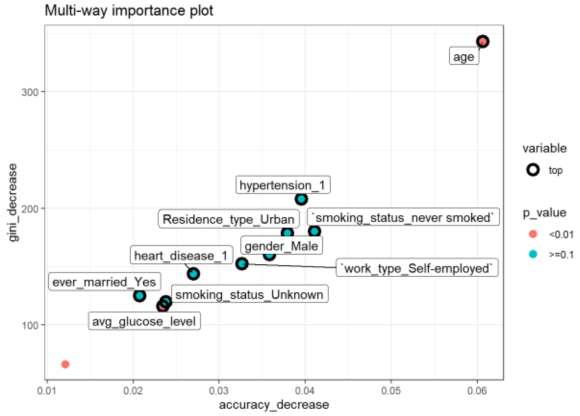
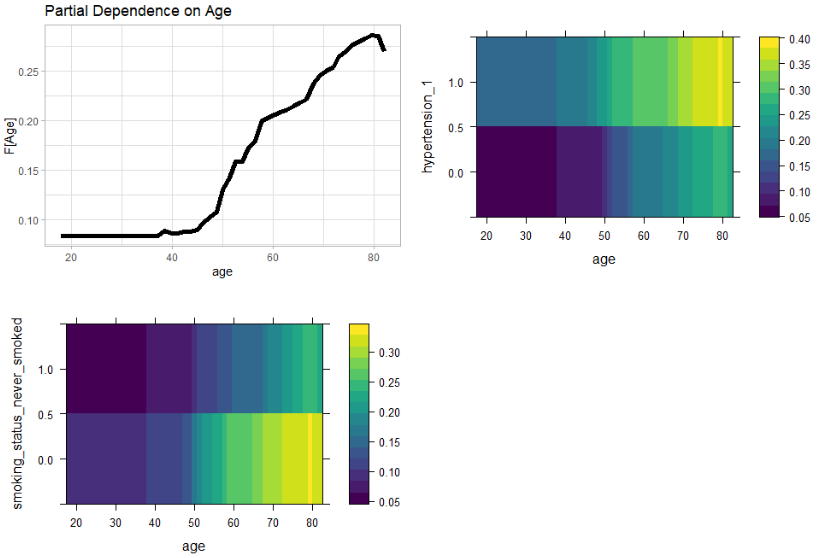


Figure 8: RandomForest – Partial Dependence Plots

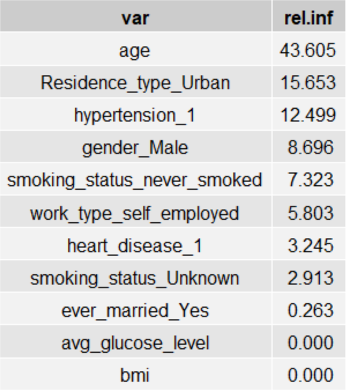


Gradient Boosting Classifier

The final model created was a gradient boosted classifier (GB). A similar procedure was used for this model as the RF model. The training set was fitted with a Cohens Kappa metric and a tune grid of n.trees = c(1000,2000), interaction.depth = c(1:5), and shrinkage = seq(0.005, 0.01, length.out = 10). A static value of 10 for the min observations in node was used. **Figure 19** shows the results of this optimization, giving us the hyperparameters shrinkage = 0.005, interaction.depth = 1, ntrees = 1000, and minobsinnode = 10.

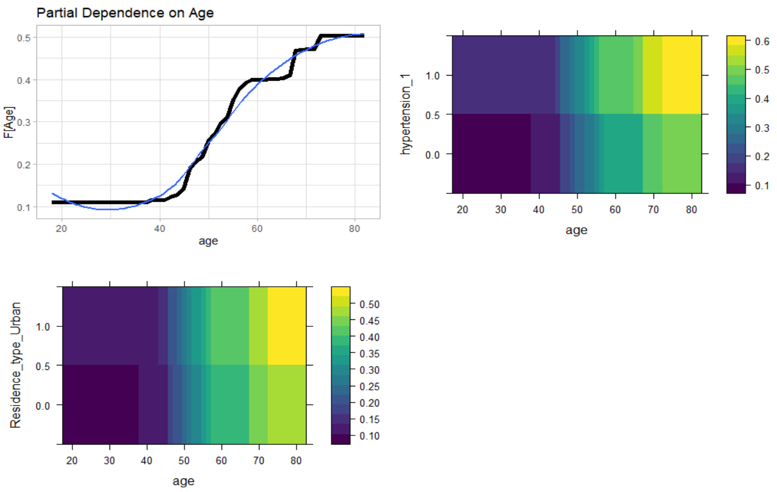
The final model was created using a 10-fold CV and luckily the output for this model comes with the relative importance of each variable used. As shown in **Figure 10**: Gradient Boost – Influence Table age continues to be the most important factor. GB disagreed slightly with the first two models on how much variables mattered but for the most part kept similar trends. Finally, the dependence plots are used to gauge impact using the top 3 variables. It's important to note that the probabilities of these graphs were not adjusted and do not represent the true probabilities of a population. That being said, this can still be used to gauge change in probability for different variables. **Figure 9** is shown below.

Figure 10: Gradient Boost – Influence Table



var	rel.inf
age	43.605
Residence_type_Urban	15.653
hypertension_1	12.499
gender_Male	8.696
smoking_status_never_smoked	7.323
work_type_self_employed	5.803
heart_disease_1	3.245
smoking_status_Unknown	2.913
ever_married_Yes	0.263
avg_glucose_level	0.000
bmi	0.000

Figure 9: Gradient Boost – Partial Dependence Plots



Model Metrics and Selections

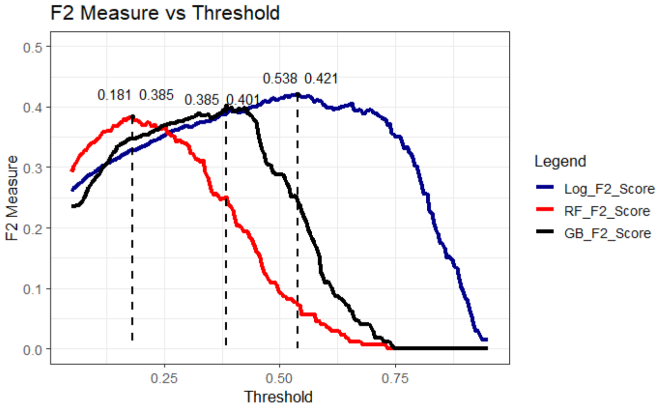
Cross-Validation Metrics

Now that the final three models are created, the goal is to pick the model that will have the best performance on our test data. Performance is subjective to the goal in mind and thus requires domain knowledge to select the best metric. In this case, false negatives (FN) are more important than false positives (FP). These models will never be used as a black and white solution to determine treatment but instead as a means of identifying negligible-risk, low-risk, and high-risk patients. Due to the high imbalance of our data, and further, of stroke cases in the population, accuracy is a bad measure of model performance. That is why the F-β score was used to optimize model threshold and determine model success. The F-β score considers recall and precision and allows for increasing the importance of the recall measure by a measure of β. In this case FNs have a huge impact and could result in a stroke for a patient deemed negligible-risk/low-risk. On the other end, a value of β too large could result in unnecessary treatment for negligible-risk/low-risk patients. The value chosen would need to be discussed with domain experts and tested against other datasets but for this application β = 2 was chosen.

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) * recall}$$

The optimized threshold was calculated using the test folds created during CV and plotted in **Figure 11** below. These values were used to create a confusion matrix against the test data that had not been seen by any of the models. The graph points to the logistic model performing the best of the 3 models but the test data will be used to select the best model. The RF model gave

Figure 11: F2 Threshold Optimizer

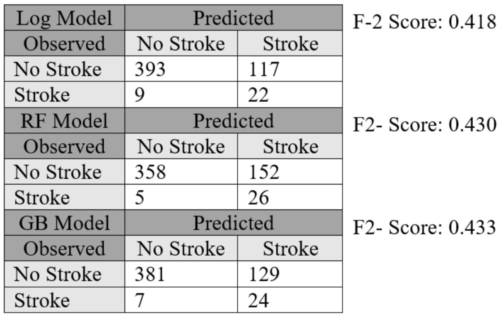


the most optimistic probabilities before being optimized, needing a threshold of 0.181 in order to maximize it's F2 score. Additional model metrics were created to get a better picture of total performance and are included in the appendix (**Figure 20**). These metrics were calculated using a threshold optimized by an F1 score. Another measure that is often used for rare cases is the PR-AUC. In our model metrics, the log model performs best as well with the GB model in 2nd place. That being said, the GB model only had 81 FNs when compared to 129 in the log model and 107 in the RF model.

Test Data Metrics and Model Selection

Using the optimized thresholds and the test data, the confusion matrix was created for each model. Using this table, we can select the GB model as the top performing one. It had the highest F2 score and although it correctly predicted 2 less strokes than the RF model, it had 23 less FPs. This model selection could change if the β is 3 or higher.

Figure 12: Final Models Confusion Matrices



Log Model	Predicted		F-2 Score: 0.418
Observed	No Stroke	Stroke	
No Stroke	393	117	
Stroke	9	22	
RF Model	Predicted		F2- Score: 0.430
Observed	No Stroke	Stroke	
No Stroke	358	152	
Stroke	5	26	
GB Model	Predicted		F2- Score: 0.433
Observed	No Stroke	Stroke	
No Stroke	381	129	
Stroke	7	24	

Conclusions and Recommendations

Looking through the process of creating these models with the data set, the following takeaways were found:

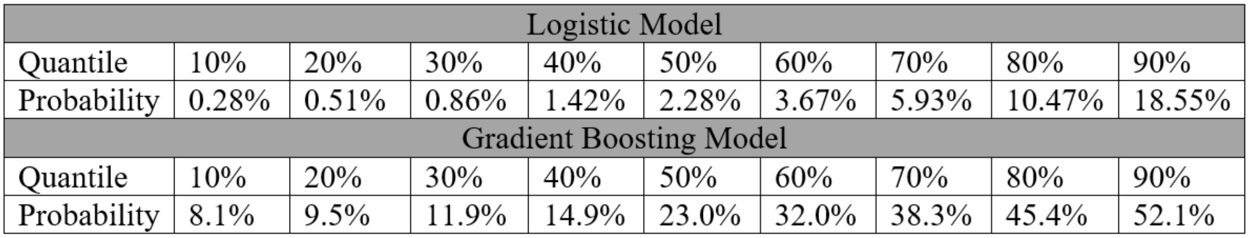
1. Age is the most impactful covariate when predicting stroke and risk increases exponentially. This is consistent with common medical knowledge
2. Gender is not indicative of increased risk when controlling for all other variables
3. BMI was an insignificant measure of stroke risk for all three models
4. Heart disease and hypertension significantly increased risk of stroke. This is consistent with common medical knowledge

5. Having never smoked decreased risk of stroke. This is likely due to other confounding variables pointing to an overall healthier lifestyle
6. Residence type was not considered statistically significant for the logistic model but still maximized the AIC metric. In all three models, living in an urban residence pointed to an increased probability of stroke.
7. Marriage status was insignificant in the logistic model and of low importance in the other two models

The overall goal of this project was met, which was to determine if social determinates had influence over the risk of stroke on a regular population. This is particularly useful since currently the widely used CHADS-VAS score is specific to patients with atrial fibrillation (afib), which means there is no model/score to stratify patients without afib for risk of stroke. That being said, it is unlikely this model can be used in any practical way as it stands. Models used for healthcare decisions have to be thoroughly vetted and often require independent studies before being recommended as a guideline for care. Additionally, the CHADS-VAS score was built on a dataset of 90,000+ patients with a medically accurate observation for each patient. The biggest downfall of this project was the data used. Because it comes from a confidential source, little insight is known about its validity, or the specifics of the measurements taken. For example, there is no mention about how the average glucose was calculated. The gold standard is to measure the bloods hemoglobin A1C to estimate average glucose levels over the past 3 months, and it must be assumed this is what the data set values for average glucose level used.

The most impactful portion of model is its ability to stratify patient risk for treatment. In **Figure 13** two quantiles are shown. The logistic model's CV stroke predictions, with adjusted probabilities, and GB model's quantiles. The stratification can be done using these quantiles and a threshold for high risk could determine by the F-2 optimization threshold. I included the logistic model because it is beneficial to be able to infer patient risk as well, even if that model

Figure 13: Model Quantiles



Logistic Model									
Quantile	10%	20%	30%	40%	50%	60%	70%	80%	90%
Probability	0.28%	0.51%	0.86%	1.42%	2.28%	3.67%	5.93%	10.47%	18.55%
Gradient Boosting Model									
Quantile	10%	20%	30%	40%	50%	60%	70%	80%	90%
Probability	8.1%	9.5%	11.9%	14.9%	23.0%	32.0%	38.3%	45.4%	52.1%

isn’t being used specifically. Using the adjusted probability threshold of 6.71%, a high-risk patient would fall in the 70th percentile in the logistic model. The probability threshold for the GB model is 38.5%, which would also fall in the 70th percentile. Using the CHADS-VAS scoring as a rule of thumb, a low-moderate risk (first level treatment is recommended) is considered at 0.6% and moderate-high risk is considered at 2.2%. This means a patient falling in the 20th percentile would be a low-moderate risk and above the 50th percentile would be a moderate-high risk. It’s important to note the probabilities seen in the logistic model do not translate perfectly to the risk shown in the CHADS-VAS score because they are observing a specifically patients with afib.

I see this as a great justification for more studies to be performed and refining the models further. The true benefit to this approach would need a higher quality data set to materialize. In addition, input from subject matter experts would be needed when deciding the best way to stratify risk.

References

[1]: Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology, 179(6), 764–774. https://doi.org/10.1093/aje/kwt312

Appendix

Figure 14: Boxplot – Age – AGL - BMI

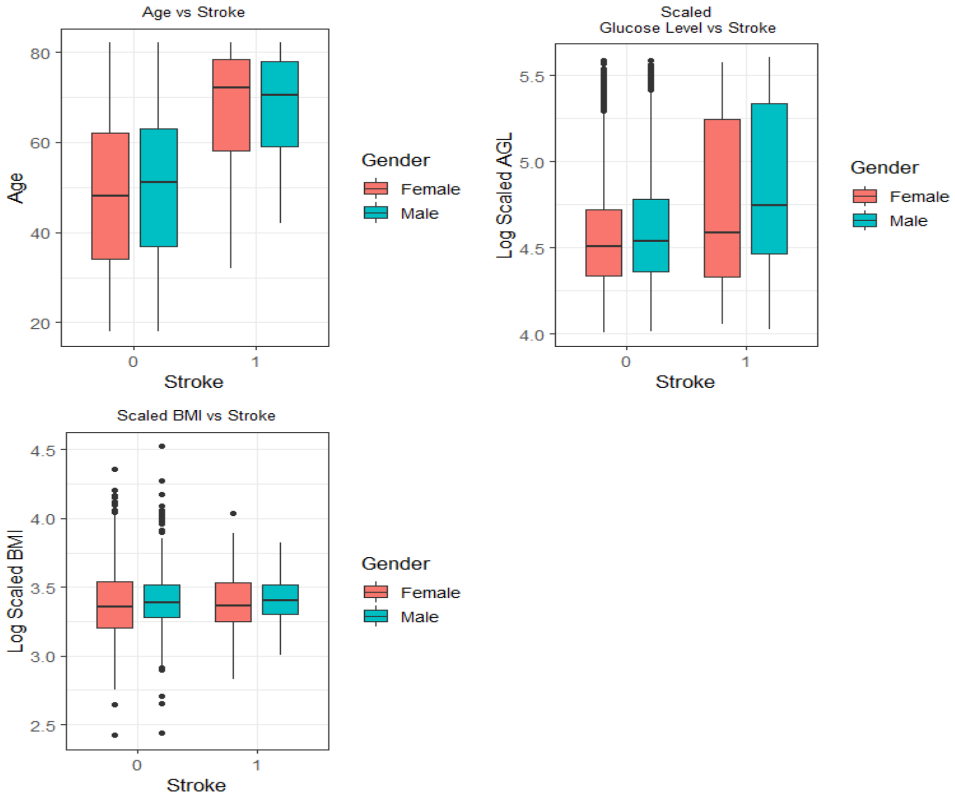


Figure 15: Pie Chart – Work Type – Residence Type – Smoking Status - Married

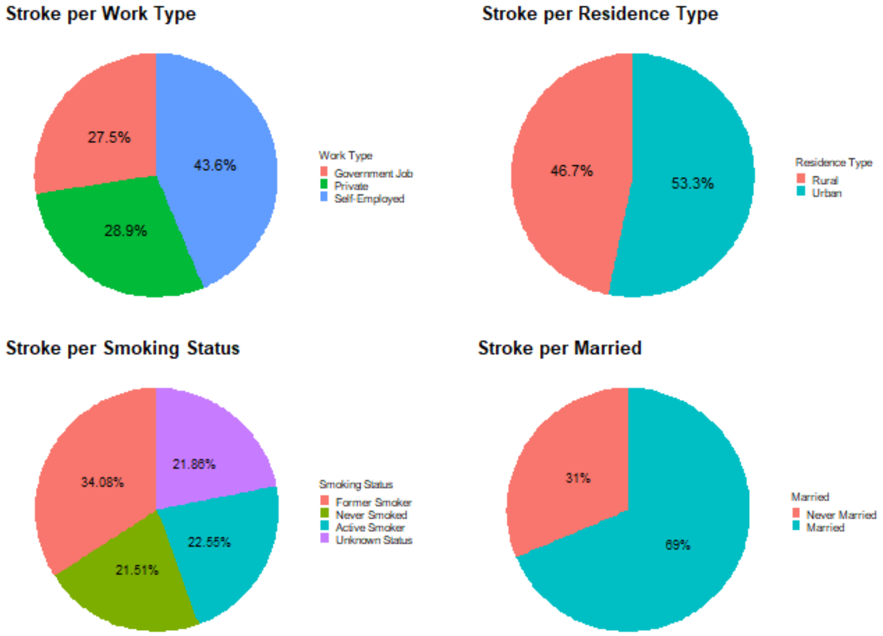


Figure 16: MICE BMI Imputation

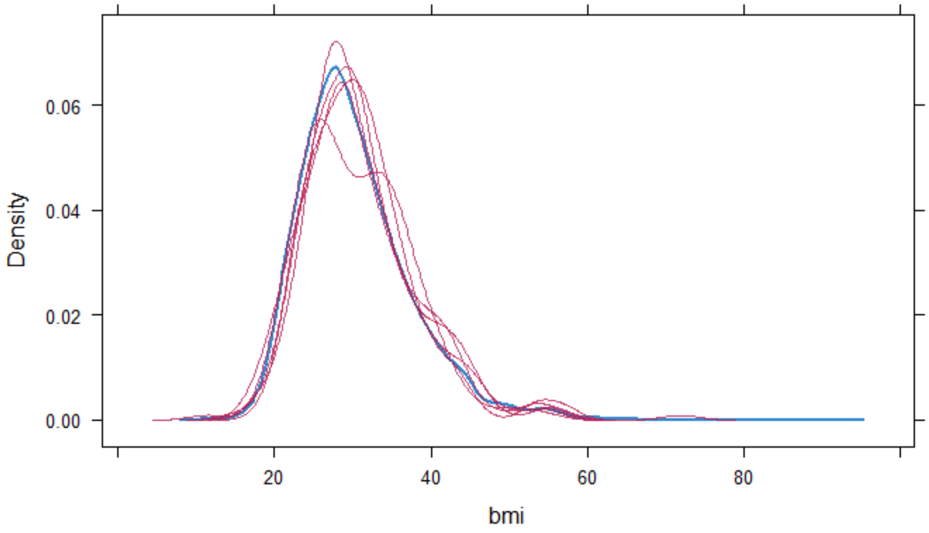
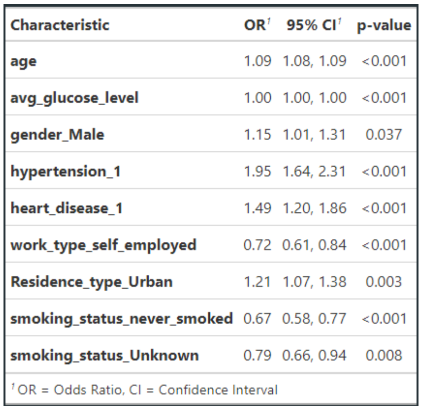


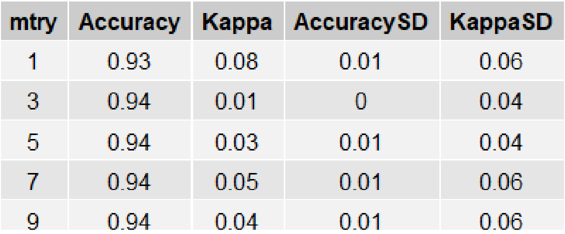
Figure 17: Temporary Log Model



Characteristic	OR[1]	95% CI[1]	p-value
age	1.09	1.08, 1.09	<0.001
avg_glucose_level	1.00	1.00, 1.00	<0.001
gender_Male	1.15	1.01, 1.31	0.037
hypertension_1	1.95	1.64, 2.31	<0.001
heart_disease_1	1.49	1.20, 1.86	<0.001
work_type_self_employed	0.72	0.61, 0.84	<0.001
Residence_type_Urban	1.21	1.07, 1.38	0.003
smoking_status_never_smoked	0.67	0.58, 0.77	<0.001
smoking_status_Unknown	0.79	0.66, 0.94	0.008

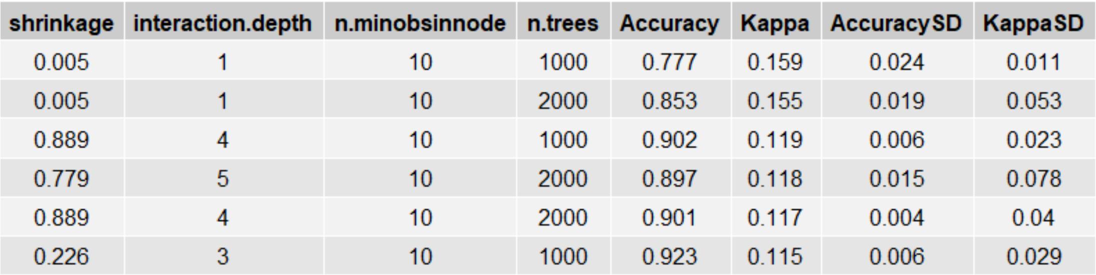
[1] OR = Odds Ratio, CI = Confidence Interval

Figure 18: RandomForest Grid Hyperparameter Optimization



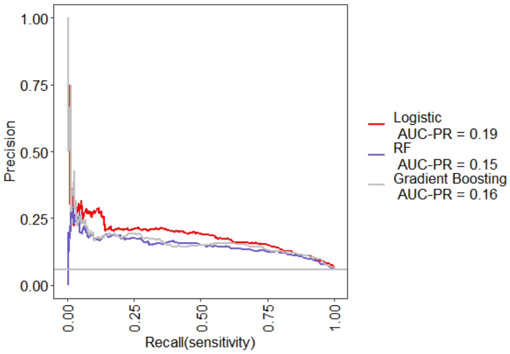
mtry	Accuracy	Kappa	AccuracySD	KappaSD
1	0.93	0.08	0.01	0.06
3	0.94	0.01	0	0.04
5	0.94	0.03	0.01	0.04
7	0.94	0.05	0.01	0.06
9	0.94	0.04	0.01	0.06

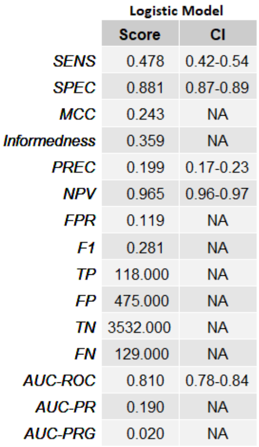
Figure 19: Gradient Boosting Grid Hyperparameter Optimization



shrinkage	interaction.depth	n.minobsinnode	n.trees	Accuracy	Kappa	AccuracySD	KappaSD
0.005	1	10	1000	0.777	0.159	0.024	0.011
0.005	1	10	2000	0.853	0.155	0.019	0.053
0.889	4	10	1000	0.902	0.119	0.006	0.023
0.779	5	10	2000	0.897	0.118	0.015	0.078
0.889	4	10	2000	0.901	0.117	0.004	0.04
0.226	3	10	1000	0.923	0.115	0.006	0.029

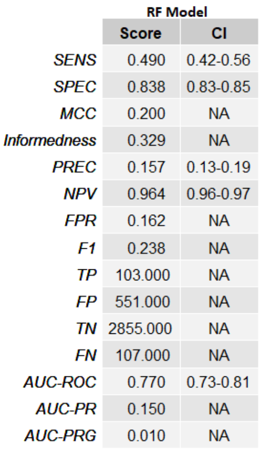
Figure 20: Final Model Metrics





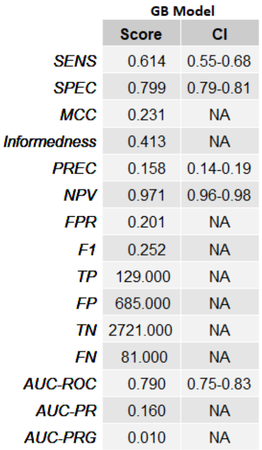
Logistic Model

	Score	CI
SENS	0.478	0.42-0.54
SPEC	0.881	0.87-0.89
MCC	0.243	NA
Informedness	0.359	NA
PREC	0.199	0.17-0.23
NPV	0.965	0.96-0.97
FPR	0.119	NA
F1	0.281	NA
TP	118.000	NA
FP	475.000	NA
TN	3532.000	NA
FN	129.000	NA
AUC-ROC	0.810	0.78-0.84
AUC-PR	0.190	NA
AUC-PRG	0.020	NA



RF Model

	Score	CI
SENS	0.490	0.42-0.56
SPEC	0.838	0.83-0.85
MCC	0.200	NA
Informedness	0.329	NA
PREC	0.157	0.13-0.19
NPV	0.964	0.96-0.97
FPR	0.162	NA
F1	0.238	NA
TP	103.000	NA
FP	551.000	NA
TN	2855.000	NA
FN	107.000	NA
AUC-ROC	0.770	0.73-0.81
AUC-PR	0.150	NA
AUC-PRG	0.010	NA



GB Model

	Score	CI
SENS	0.614	0.55-0.68
SPEC	0.799	0.79-0.81
MCC	0.231	NA
Informedness	0.413	NA
PREC	0.158	0.14-0.19
NPV	0.971	0.96-0.98
FPR	0.201	NA
F1	0.252	NA
TP	129.000	NA
FP	685.000	NA
TN	2721.000	NA
FN	81.000	NA
AUC-ROC	0.790	0.75-0.83
AUC-PR	0.160	NA
AUC-PRG	0.010	NA