Introduction

Strokes are a leading cause of death worldwide and are a source for major disability in adults. For this analysis, I plan to utilize a dataset on strokes from Kaggle to fit artificial neural network (ANN) and random forest (RF) models, to predict if a stroke would occur based on various predictors. The goal of this analysis is to find a robust model which is optimized for predicting stroke patients. The intention of this study is for the results to be utilized by medical practitioners who have a particular focus area in strokes, such as neurologists, to save lives.

Data Cleaning and Exploratory Analysis

To start, the Kaggle dataset on strokes was loaded into R, which contained 5110 rows and 12 columns. The columns in the dataset were as follows: “id”, “gender”, “age”, “hypertension”, “heart disease”, “ever_married”, “work_type”, “Residence_Type”, “avg_glucose-level”, “bmi”, “smoking_status”, and “stroke”. The “id” column was removed since it wasn’t needed for the analysis, and all columns except for “age”, “bmi”, and “avg_glucose-level” were converted to factors. Preliminary analysis noted the dataset was imbalanced, there were only 249 entries for stroke patients which corresponded to about 5% of the overall dataset. Furthermore, the BMI column had 201 entries with “N/A” values. First, it was considered to remove these entries, but if removed the new dataset would only have about 4.1% of entries be stroke patients. Given this dataset was already very imbalanced, it was decided the next appropriate step was to impute the missing values. To decide if median, mean, or kNN imputation should be used, a normal probability plot was created for BMI with results in Figure 1 below:

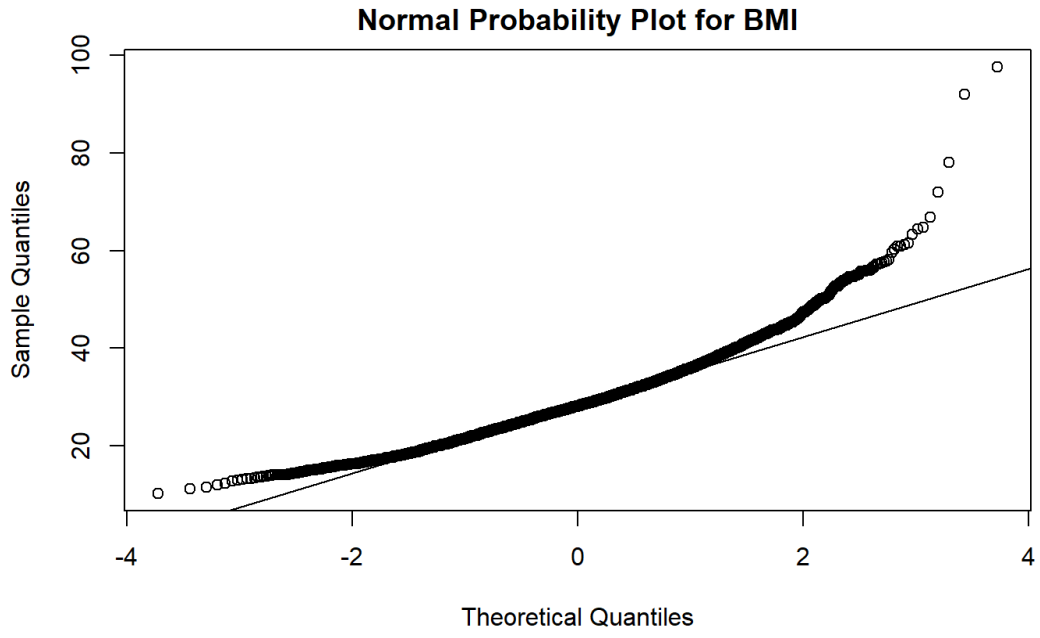

Figure 1: Normal probability plot for available BMI data

Given Figure 1's skewness, it was decided kNN imputation would be used for the missing BMI values. Feature engineering also revealed the "other" level in "gender", and "never_worked" level in "work_type" should be removed, since these observations made up less than 0.5% of the total dataset. This was done to simplify a very complex dataset, with the goal to provide an optimal input for the ANN and RF models. Furthermore, since BMI and average glucose level each exhibited right skewness, each were log transformed to keep data as consistent as possible for further analysis. Finally, the correlation plot between the 3 numerical predictors of "age", "log_bmi", and "log_avg_glucose_level" only revealed a moderately strong positive correlation between "age" and "log_bmi" predictors.

Now that the data was cleaned, exploratory data analysis was done to find if any predictors exhibited a relationship with the response. Density curves with fill by the response were created for numerical predictors, while proportional bar charts by response were created for factor predictors. The following figures below are examples of predictors which exhibited an interesting relationship with the response.

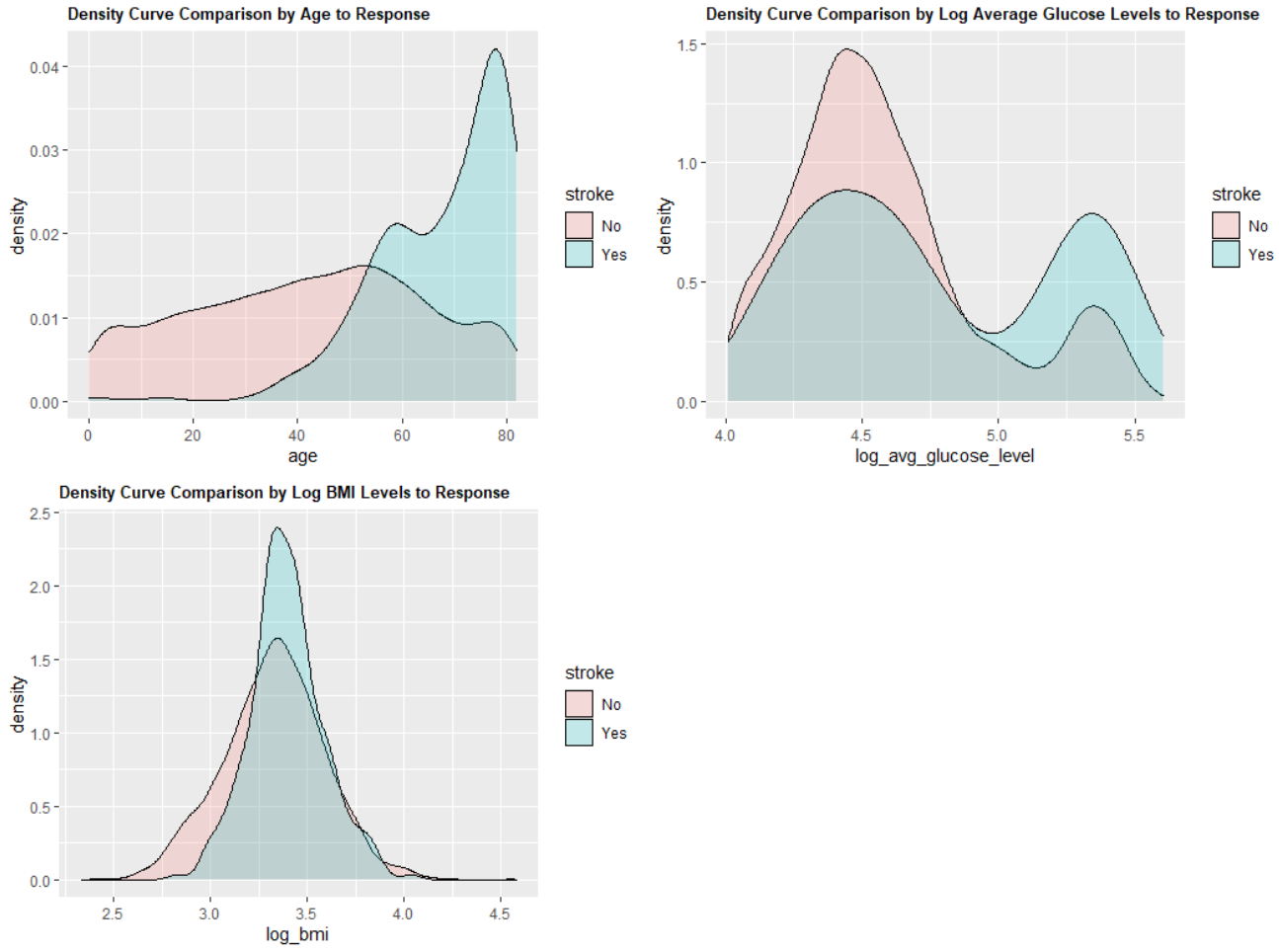

Figure 2: Density Curves for Age, Log Average Glucose Level, and Log BMI by Response

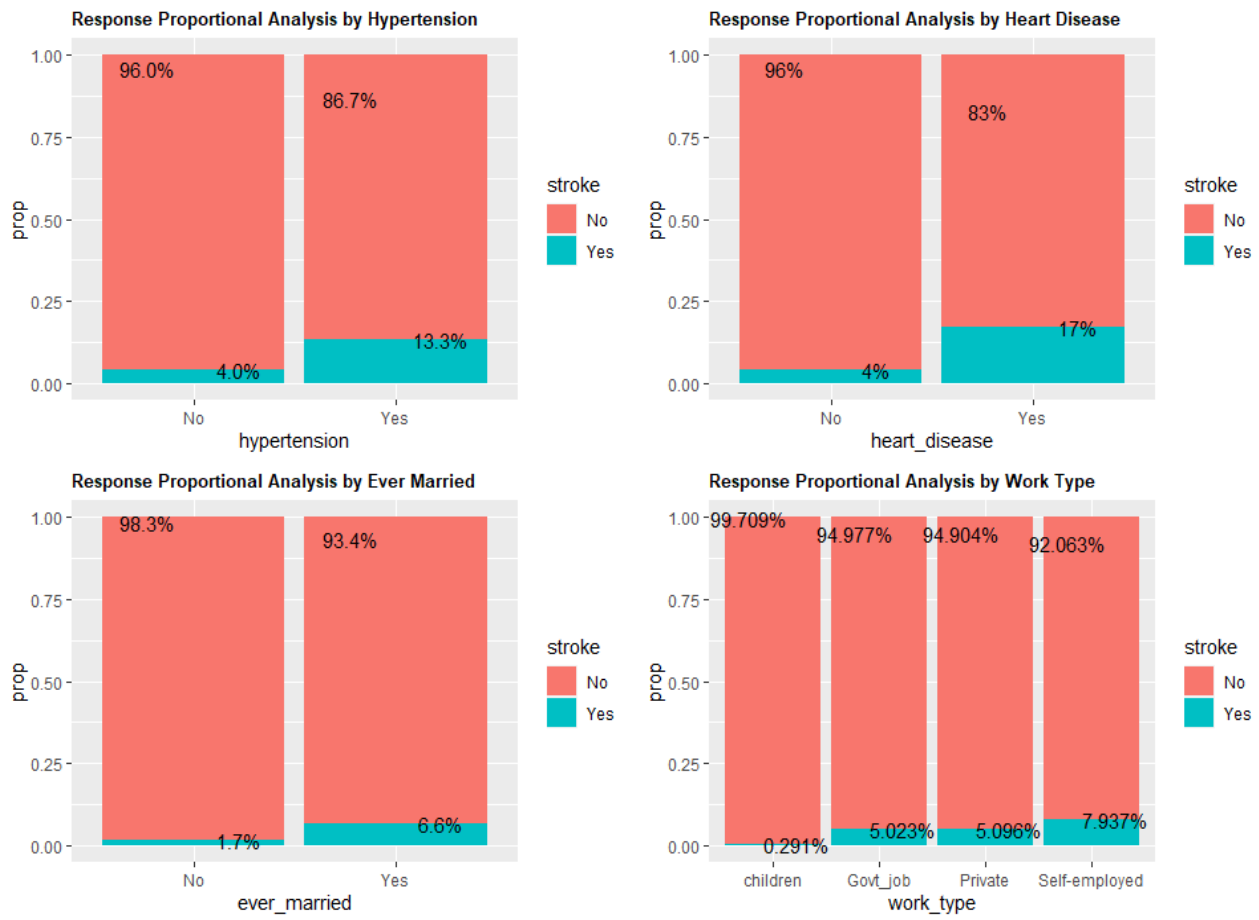

Figure 3: Proportional Bar Charts for Hypertension, Heart Disease, Ever Married, and Work Type by Response

Model Fitting and Selection Process

RF and ANN models were each chosen since it was expected this would be a complex, non-parametric response. Given the moderately large dataset size of over 5000 entries, the hope was this would be enough data to train an ANN model for the complex response. For the RF model, since a subset of predictors exhibited more of a relationship with the response than others, the hope was this would be optimal for the complex response. All predictors will be used to fit each model. Furthermore, each model had the following tuning parameter ranges while trained in caret:

- For random forest, the initial tuning parameter range was chosen to be 1 to 10 since the final parameter would be in that range. 1 would represent the simplest model, while 10 would represent the most complex since there are 10 predictors in the dataset.
- For ANN, the parameter being tuned was weight decay, while hidden node size was fixed at 1. The weight decay range chosen was from 10^-7 all the way to 10^6. The lower parameter was chosen since weight decay parameters are often smaller for ANN modeling as compared to ridge regression. The higher parameter was chosen since this dataset was relatively large, and based on testing with the metric below.

Now with the models defined, the next challenge was to find an optimal metric to train and assess each model on, since this dataset exhibited class imbalance. A novel, weighted brier score was used to both train and test each model. This metric was essentially a negative, weighted mean version of the brier score, with a weight of 2 assigned for stroke patients, and a weight of 1 for non-stroke patients. The idea with the weight was it would serve as an extra penalty if a square error was very high for a stroke patient. This fits our goal to optimize a model for predicting stroke patients.

After the model fitting process, it was found random forest was clearly the optimal model, with a metric consistently around -0.078 for all training data. The ANN models had a metric around -0.25 and converged for all available training data. The average metric on the model honest assessment across the 5-fold double CV was also -0.078, with a high of -0.090 for fold 4, and a low of -0.068 for fold 5. The random forest model with a tuning parameter of 8 was chosen to be the final model. This is based on 8 appearing once from fold 5 of the double CV, and from the 5-fold inner CV on the entire dataset. The metric by tuning parameter range is presented for the final model in Figure 4 below.

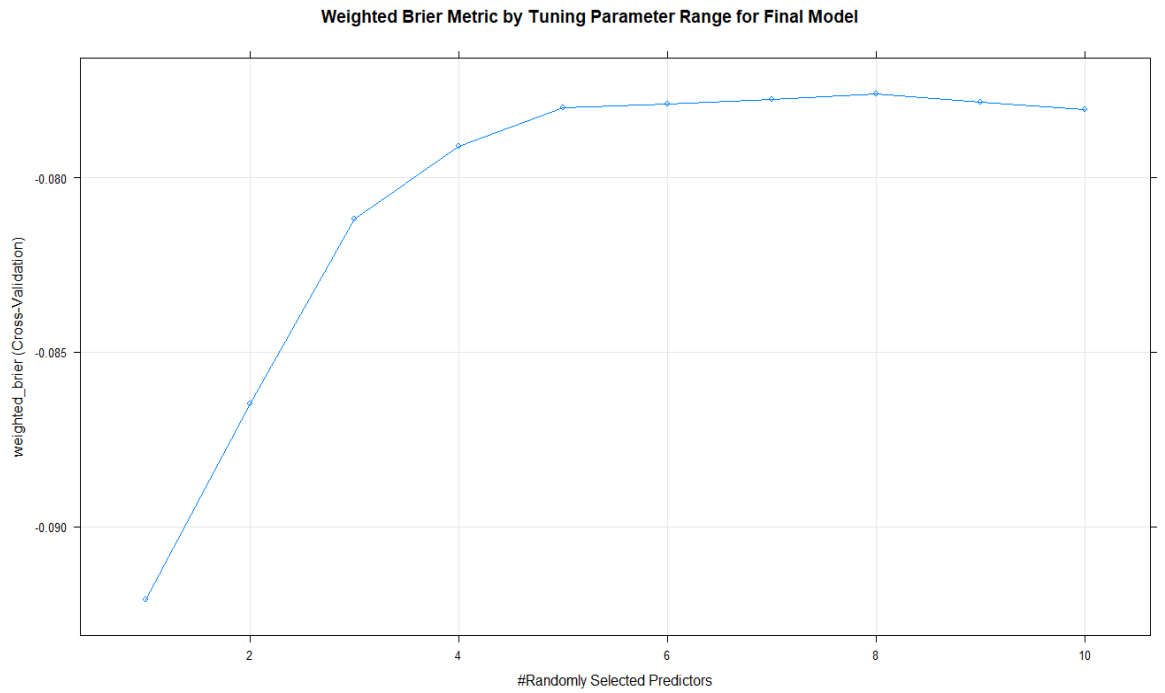

Figure 4: Optimal Tuning Parameter for Metric for Final Random Forest Model.

Two Most Important Predictors

The two most important predictors from the final random forest model are presented in Figure 5 below.

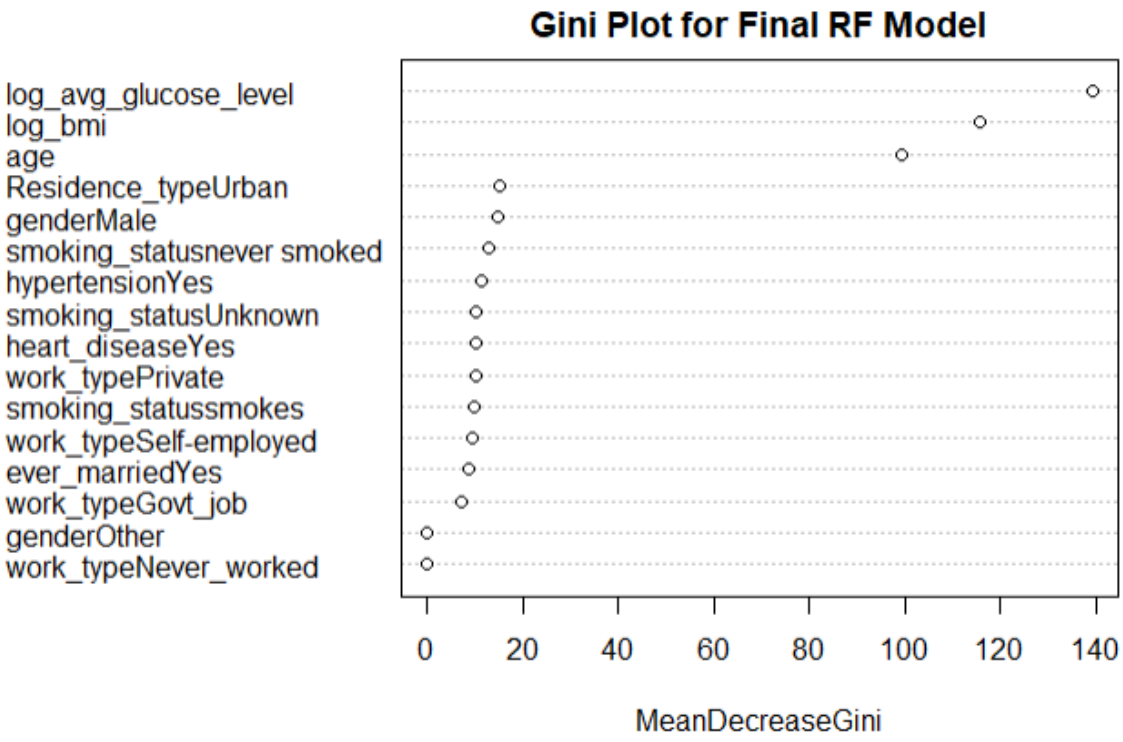

Figure 5: Gini Plot for Final Random Forest Model

From Figure 5, and caret's variable importance, the two most important predictors were clearly "log_avg_glucose_level" and "log_bmi." This makes sense based on my knowledge of strokes, as these two predictors are key health indicators. There are numerous online research publications which state how higher levels for each predictor correspond to a higher risk of stroke. Figure 2 above also supports each predictor being positively related to the response. Furthermore, the mean of each predictor is higher for stroke patients, as compared to non-stroke patients in this dataset. Figures 6 and 7 were created below to visualize how the final RF model is predicting the probability of stroke rises by "log_avg_glucose_level" and "log_bmi" levels.

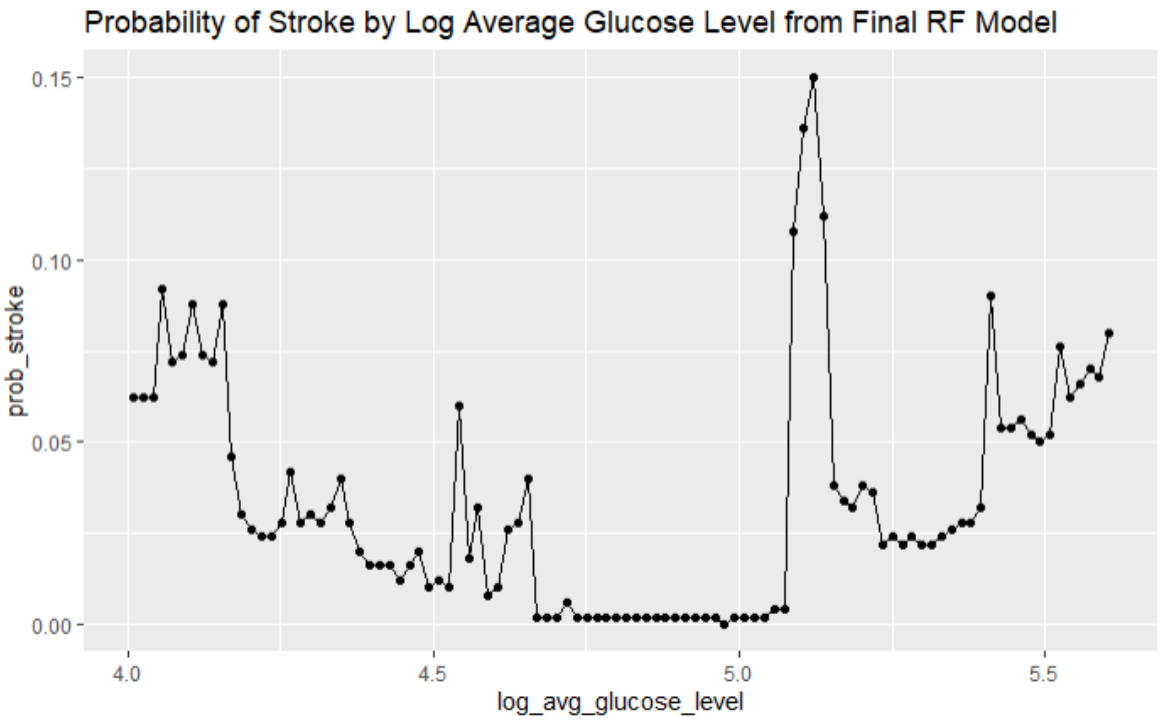

Figure 6: Probability of Stroke by Log Average Glucose Level from Final RF Model

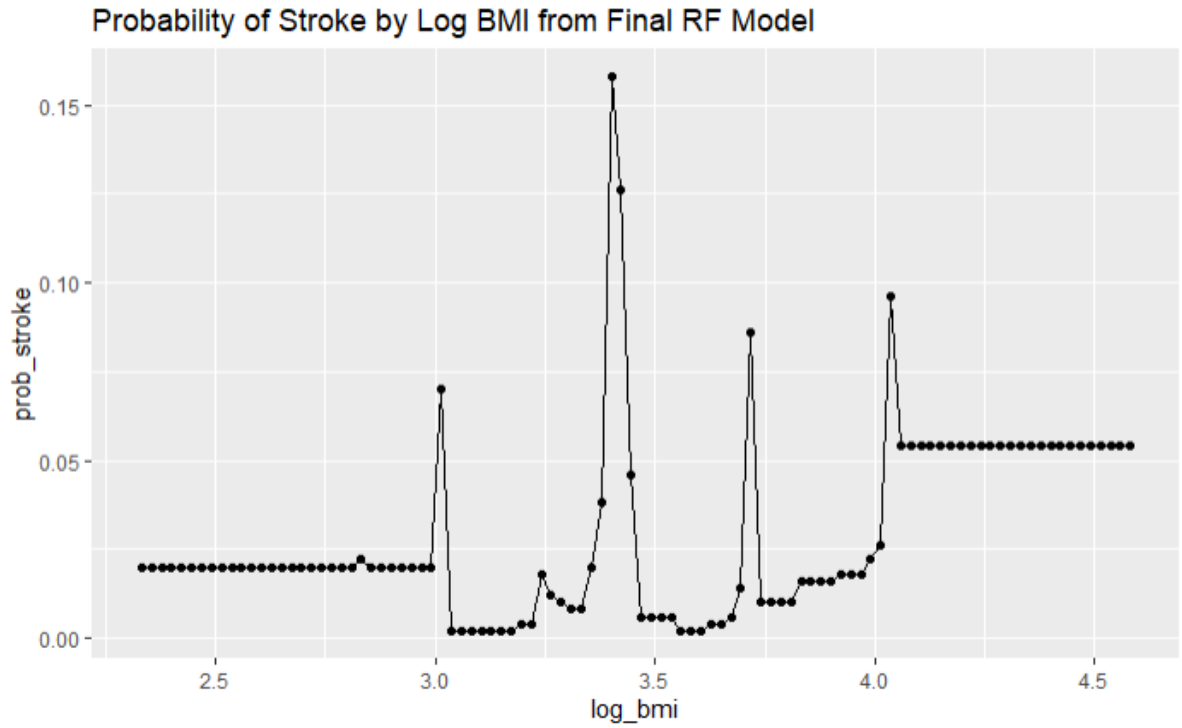

Figure 7: Probability of Stroke by Log Average Glucose Level from Final RF Model

These figures also support higher levels of glucose and BMI lead to a higher chance of stroke. The probability spikes in each figure also make sense, given the mean "log_bmi" and mean "log_avg_glucose_level" for stroke patients are nearly 3.4 and 4.8, respectively. The medical community must continue to educate the public on the increased stroke risk for higher BMI and average glucose levels to promote healthier lifestyle choices.

Model Limitations

While the intention of this model fitting process was to be robust, improvements may be possible. First, there may be a more optimal weight ratio for the weighted brier metric for stroke verse non-stroke patients. As a test, another model was fit from the random forest package with a randomly sampled training dataset that was 2/3rd of the overall dataset. The remaining 1/3rd of the dataset was used as a test dataset. When testing this model with a default tuning parameter of 3, the average probability of "yes" for stroke observations in the test dataset was 0.138. When re-fitting this model with a tuning parameter of 8, the average probability of "yes" for stroke observations in the test dataset was 0.167. This indicates good progress, the goal of our modeling is to optimize prediction for stroke patients, which means we would like the probability of "yes" for stroke observations to be as high as reasonably possible. However, it was also noted that re-fitting this model with a tuning parameter of 10 resulted in this probability to be 0.170 for the same test observations. Furthermore, other metrics exist for class imbalance problems such as AOC, Kappa, PRC, and the original Brier score. These metrics were also explored in this project, and it was noted it was a particularly difficult challenge in objectively assessing what the best metric may be for this class imbalance problem.

This project also presented challenges with the ANN model. Due to limitations in computational power, the hidden node size for the ANN model was not able to be tuned. The decay parameter was chosen to be tuned based on results it reduced the weighted brier metric much greater than tuning hidden node size. However, it was noted the better result with decay parameter was misleading. The decay parameter, which gave the best weighted brier score at -0.25, was 10^6. This parameter is exceptionally large for an ANN model. The resulting weights from this model were practically 0, and the output of the model had constant probabilities for strokes, -0.49 for "yes" and -0.51 for "no". This final model had little predictive ability and is another example of the difficulty tuning ANN models. One promising alternative model to ANN for this problem may be boosted trees, although there may also be challenges with the number of parameters needing to be tuned for a dataset of this size. A SVM model should also be explored with a combination from the "log_avg_glucose_level", "log_bmi", and age predictors.

Conclusion

In conclusion, the random forest model with the novel, weighted brier metric, proved to be an effective start for this difficult class imbalance problem with the stroke's dataset. Future research should focus on the optimal metrics, training strategies for ANN models, and other potential optimal model types for class imbalanced medical disease problems.