University of Virginia

Final Project

Foundations of Machine Learning

Professor Johnson

December 7, 2023

Summary

Our primary objective was to develop a model that predicted the likelihood a person would have a stroke using a dataset with various health-related variables. The key variables we chose to focus on were age, hypertension, heart disease, body mass index (BMI), blood sugar level, age, residential area, and employment type. Prior to building the model the data was cleaned, where missing values, specifically for the 'bmi' variable, were replaced with its average. Other variables that had errors with their values, such as 'age,' where there were in-between values, were rounded to the nearest whole number. A model was then created using a linear regression model with polynomial features and a decision tree model, combined using different methods to create the most accurate predictive model. Separately, both had higher RMSE values than together; the linear model had a value of .206, and the decision tree model had a value of .234. Initially, simple averaging was used for the predictions of the decision three and linear model, however, the RMSE value was not lower than the linear model value with .228. Then, weighted averaging was used, with the linear model weighing more than the decision tree since the RMSE value was lower. This too, however, resulted in a similar RMSE value at .226. A different way to combine them was used: residual fitting. Using the linear model predictions, along with the initial decision tree predictions, the residual values were calculated and fitted using another decision tree model. Averaging the linear and decision tree values again, this time with the predictions from the residual model, a more predictive model was achieved with a lower RMSE of .195 and r-squared value of .175, meaning enhanced accuracy of stroke prediction for the dataset, but limited variability.

Data

The data to build the final model came from two sets: the training and testing data. As described by their respective labels, the training data set was used to train the models, and the testing data set to test the model of choice. The data sets contained the same variables along with their data. The key variables used in the analysis were:

- age: Patient age, numeric
- avg_glucose_level: Blood sugar levels, numeric
- bmi: Body mass index, numeric
- heart_disease: Heart disease status, dummy (0/1)
- hypertension: Has hypertension, dummy (0/1)
- stroke: Whether or not they suffered from a stroke during the sampling period
- residence_type: Residential area, dummy (Urban, Rural)
- smoking_status: Smoking status, categorical, (Former, never, or current smoker)
- work_type: employment status/type of job, categorical, (govt_job, Self-employed, Private, children, Never_worked)

The identification variable (id), marriage status (ever_married), and gender variables were not used to build the model. As for data wrangling, only two of the variables used in the model needed to be cleaned. The 'age' variable had many floats instead of whole integer values. Decimals within the dataset were rounded up or down depending on the value. The 'bmi' variable contained a NaN value which was replaced with the mean for the variable (28.9). No normalization or scaling was necessary for the data.

When building the model, the categorical variables, work_type, residence_type, smoking_status, were one-hot encoded using the pd.get_dummies() function to encode them into

binary variables. Numeric variables, age, avg_glucose_level, and BMI were indexed using the .loc[:, function. The same transformations were performed on the variables in both the testing and training sets.

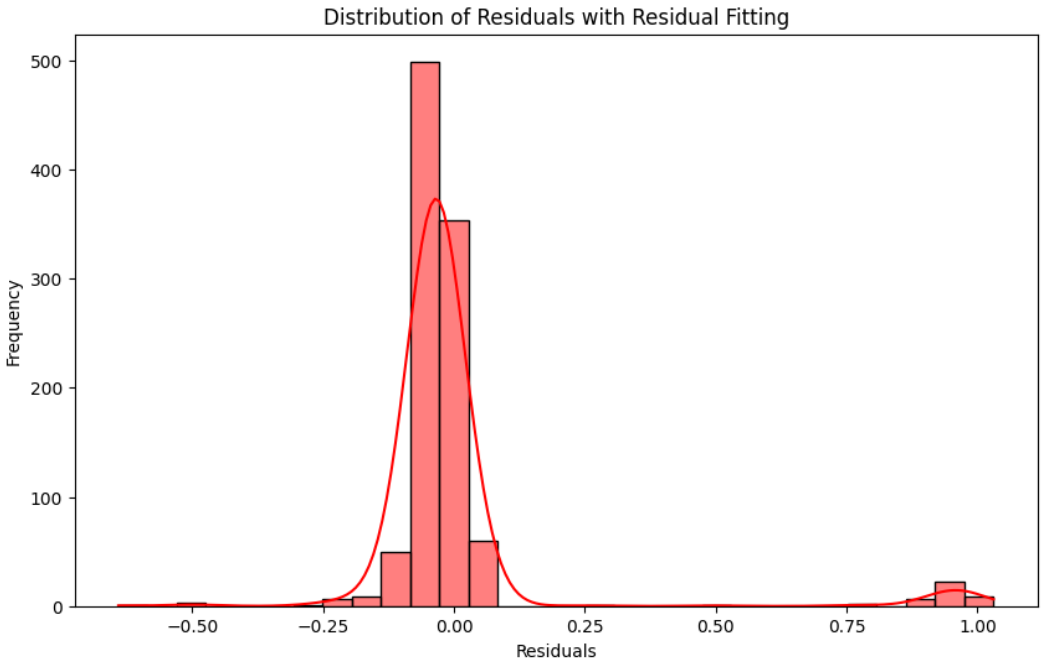
Results

Combining a linear model with polynomial features with decision trees using residual fitting, resulted in a lower RMSE value of 0.195 and r-squared value of .175. After one-hot encoding of the categorical variables and indexing the numeric ones, a linear model with polynomial features was formed. Using prior code, the RMSE value of the linear model was 0.206. Deciding to add a decision tree model in combination with the initial linear model was likely to add a better predictive performance as tested than either one tested alone. The issue arises when combining. At first, a more simplistic method was used: simple averaging. Using simple averaging to combine the predictions of both models ultimately resulted in an RMSE value of 0.227 which was higher and less accurate than the linear model with polynomial features but lower and more accurate than the decision tree model alone with a RMSE value of 0.234. Another method, weighted averaging was used. Because the linear model had a slightly better predictive ability, instead of weighing both models equally, the linear model was given a higher weight (0.6) and the decision tree model had a lower weight (0.4). This ultimately resulted in an RMSE value that was only slightly 0.226. When weighing the linear model more using different values (0.7, 0.8), and the decision tree value less (0.3 and 0.2) RMSE value stayed the same.

The apparent most efficient way of combining both models that resulted in the lowest RMSE value was combining through residual fitting, or a more proper term, “residual stacking.”

This technique uses the residuals from two separate, distinct models and is able to enhance the accuracy of predictions on a particular set by adjusting them based on the residual errors. Using the residuals from both models, a third model was fitted solely on the residuals using decision trees by averaging the linear predictions and the decision tree predictions and subtracting them from the actual target values. After residual fitting and predicting on the testing data, the predicted residuals were incorporated into the linear and decision tree combined model again via averaging and adding the new residual predictions, allowing for a more accurate predictive model. Below is a visualization of the distribution of residuals for the model using residual fitting. As you can see, there is a somewhat symmetrical distribution of residuals around zero and a higher frequency of residuals closer to zero, suggesting that the predictions the model makes are typically accurate.

Figure 1



Conclusion

In conclusion, health-related and outside variables such as age, hypertension, heart disease, body mass index (BMI), blood sugar level, residential area, and employment type do seem to have some impact on someone's stroke likelihood. However, these are not the only factors that can impact whether or not a person may have a stroke. Our model, which appears to have good predictive ability with an RMSE value of .195, has an r-squared value is .175 which indicates that only a small percentage (17.5%) of the variability that is associated with stroke likelihood can be explained by our model and these variables.

Like any health-related issue, there are a multitude of factors besides the variables listed that can impact one's health conditions. In terms of stroke likelihood, other lifestyle and environmental factors as well as prior health conditions besides heart disease or hypertension are just as likely to play a role as any of the other variables listed. One variable that was not included in this data set, race, plays a crucial role in someone's stroke likelihood. According to the U.S. Department of Health and Human Services (Office of Minority Health), African-Americans are 50% more likely to have a stroke compared to their white counterparts. African-American women are twice as likely to have a stroke than non-Hispanic white women and African-American men are 70% more likely to die from a stroke than non-Hispanic white people.

In summary, although our model does not capture every nuance of stroke likelihood, it does give great insight into factors that may be incredibly important. The scope of our model is limited to very few variables but is a great starting point. In the future, it is incredibly important to add a multitude of different factors as understanding the main determinants of stroke likelihood is bound to be beneficial to many communities.

Works Cited

"Stroke and African Americans." *Office of Minority Health*, minorityhealth.hhs.gov/stroke-and-african-americans#:~:text=Black%20men%20are%2070%20percent,to%20non%2DHispanic%20white%20women. Accessed 7 Dec. 2023.