University of Virginia

Project 2 Paper

Group 12

Sydney Stokes, Alijah Wyatt, Andrea Nava

Foundations of Machine Learning

Professor Johnson

December 8th, 2023

Summary

Our objective was to develop a model that predicted the likelihood a person would have a stroke using a dataset with various health-related variables, such as age, bmi, or smoking status. The key variables we focused on were age, whether or not they had hypertension, whether or not they had a heart disease, bmi, and their average glucose level. We first started by cleaning the data, where we handled the missing values, replaced specific values with whole numbers, and replaced NaN values with the average. After cleaning the data, we began by creating a combined model of the decision tree regressor and a linear regression model with polynomial features, using the weighted average. We transformed the training dataset, creating dummy variables for the categorical features. However, the resulted RMSE value was 0.23, which meant the model was rather weak in predicting the chances of a stroke occurring. Next, we created a new combined model, but this time using regular averaging. The resulted RMSE value was a 0.23, which meant it also had a weak ability to predict the chances of a stroke occurring. The third combined model we created used residual fitting instead. This third combined model of a decision tree regressor and a linear regression model achieved a combined RMSE of approximately 0.195. Since the alternative combined models using regular averaging and weighted averaging both resulted in a higher combined RMSE value, this indicates that the combined model with residual fitting performed better and this approach enhances predictive accuracy. This analysis demonstrates the effectiveness of a combined model utilizing both linear and tree-based models, with an additional layer of complexity introduced through residual fitting, leading to improved predictive performance in the occurence of a stroke.

Data

For our project, we used the two datasets provided to us, the training dataset and the testing dataset, each containing several variables. The variables in the datasets ranged from lifestyle choices and demographics to factors analyzed from an individual's medical history. The numeric variables are the patient's age, their blood sugar levels, their body mass index, their study identification number, whether they have a heart disease, and whether they have hypertension. The categorical variables are their residence type, their work type, if they were ever married, and their smoking status. However, the primary variable we used was the stroke variable. It represented the dependent variable in our predictive models.

Results

Conclusion