Summary

Throughout Project 2, Group 22 built algorithms to predict the likelihood a person has a stroke, utilizing the variables and data accessed from the provided datasets. To do this, the group first prioritized cleaning the data; rounding up numeric values, type casting variables, handling missing values, and securing the unique inputs of several categorical variables. After the data cleaning, several visualizations were made to illustrate data trends and insights. Particularly, the first bunch of visualizations revealed the counts for the unique values of the testing dataset, with the second bunch doing the same for the training dataset. Afterward, data modeling was conducted. First, a linear regression model was made between the stroke variable and the categorical variables, producing root mean squared error and R-squared values of approximately 0.2059 and 0.0849, respectively. Secondly, a linear regression model between the stroke variable and the numeric variables with polynomial expansion was made, ultimately showing that the highest R-squared value produced by any of the polynomial expanded models was roughly 0.0863, which was achieved with a degree of 2. Followingly, a regression tree was made for the stroke variable, which discovered the best model to have a depth of 2 an R-squared value of roughly 0.0635, and a root mean squared error of approximately 0.1867. A classification tree was then made for the stroke variable, ultimately revealing the best model having a depth of 5 and an accuracy of roughly 0.9419. Followingly, a logistic regression was used, root mean squared error of roughly 0.2211 and an R-squared value of approximately -0.0514. The last predictive algorithm assessed for the stroke variable was the KNN classification model, which ultimately also revealed a root mean squared error of 0.2211. While the group explored a variety of predictive models to predict the chance of someone having a stroke, most of these algorithms produced relatively weak correlations between the stroke variable and other data values, portrayed by low R-squared and high RMSE values. The best model is the linear regression model with polynomial expansion of degree 2, with an R-squared value of approximately 0.0863. But as the algorithm encompasses the highest R-squared value amongst the rest, it is still quite low, as it suggests that only approximately 8.63% of the variability in the dependent variable, stroke, is explained by the model.

Conclusion

For Project 2, Group 22 undertook a comprehensive analysis to predict the likelihood of a person having a stroke, leveraging various algorithms upon the provided testing and training datasets. Before the modeling, the data was cleaned to ensure the integrity of the datasets, and several visualizations were made to present insights and trends gathered from the variables. Throughout the project, 6 predictive models were used: linear regression, linear regression with polynomial expansion, regression tree, classification tree, logistic regression, and a KNN classification model, each illustrating distinct methods to evaluate the correlations between the stroke variable and other variables. While still low, the best predictive model exercised was the linear regression model with polynomial expansion of degree 2, producing an R-squared value of approximately 0.0863. This indicates that approximately 8.63% of the variability in the dependent variable, stroke, is explained by the model, thus suggesting a very modest explanatory power.
The main source of criticism many may highlight is the relatively low R-squared values acquired across various models. After all, if the goal of the project is to build algorithms to predict the likelihood of a person having a stroke, it is understandable why it may be worrisome to receive a rather low indicative measure for forecasting the incidence of this medical emergency. Because of the low R-squared values obtained, many would go as far as to say that the project was a complete failure! This is far from accurate.

The acknowledgment of the weak correlations between the stroke variable and other data values demonstrates transparency in the analysis, recognizing the inherent complexities of the dataset and the limitations of the predictive models exercised. More specifically, the predictive models reveal that the provided datasets have a weak explanatory power in predicting the incidence of a stroke. While this might be disappointing, it still provides relevant information, revealing that the variables Group 22 used to form algorithms provide a weak predictive power for someone experiencing this medical emergency. The project's merit lies in its comprehensive exploration of diverse models, contributing to a nuanced understanding of the data, and ultimately revealing the consistent modest predictive power gathered across numerous models. Despite the challenges reflected in the relatively low R-squared and high RMSE values, these findings offer valuable insights and provide a foundation for further exploration. It pushes researchers to continue the exploration of better datasets to hopefully provide better explanatory strength for strokes.

If this project were to be reproduced in the future, many steps could be taken to improve the quality of its results and conclusions. The primary way to do so would be by incorporating a larger amount and more advanced techniques to enhance the predictive power of the models. For example, through the application of ensemble methods, such as random forests or gradient boosting, researchers could combine the strengths of multiple models to improve their collective predictive performance. Another advanced algorithm that could be implemented is a temporal statistical analysis, enabling future researchers to assess and model the behavior variables across distinct periods. Lastly, through the inclusion of external datasets containing relevant information not present in the provided datasets, a more holistic understanding of the factors influencing stroke occurrence could be attained. By incorporating these methods outside of the scope of this project, much advancement can be achieved.