STROKE PREDICTION ANALYSIS
CS-688

Justin Shin, Darryl Anthony, Alex Bowman

I. ABSTRACT

With around 795,000 cases every year, Stroke is a leading cause of death and disability in the United States. Stroke is a preventable medical condition which makes analyzing the underlying causes even more worthwhile. There are multiple risk factors associated with Stroke and a good majority of it can be limited by altering lifestyle. This article focused on exploring the correlation between lifestyle and possibility of having a stroke. The dataset we used included both medical features as well as lifestyle related factors. Our methodology included using two baseline models – Logistic Regression and Decision Tree, and a Neural Network. The goal was to implement our model and use it on the whole dataset, and compare it to a second dataset which was essentially the original dataset but with lifestyle related features removed. Our analysis showed that for both baseline models, the accuracy of stroke prediction dropped when lifestyle features were removed, whereas the Neural Network failed to register a difference in accuracy. For Logistic Regression, our model exhibited around 4% more false negatives when lifestyle removed. On the other hand, our Decision Tree model showed no false negatives when lifestyle features were included. Similar to Logistic Regression, our Neural Network showed an increase of around 55% in False Negatives when lifestyle features were removed. Using the results generated from our models, we concluded that lifestyle has a positive correlation with chances of stroke.

II. INTRODUCTION

Stroke is a serious medical condition and is a leading cause of death and disability in the United States. It's related to the blood arteries going to and from the brain. When the arteries carrying the blood to the brain gets blocked or ruptured, the brain is deprived of the oxygen it needs and hence the brain cells die causing a stroke. Common symptoms of Stroke include sudden numbness in the face, arm, or one side of the body. Other symptoms include difficulty talking, understanding speech, along with loss of coordination and balance. According to the CDC, around 750,000 people have a stroke every year in the United States. Of the reported cases, over 17% of these incidents result in death with around 20% of these being recurring events. Stroke is a preventable disease and there are multiple treatable risk factors.

In recent years, several studies have investigated the relationship between lifestyle factors and possibility of stroke. A recent study by Krittanawong and Kumar examined the relationship between self-employment and cardiovascular risk among the general

population in the United States [2]. The study consisted of around 30,000 patients, and after adjusting for health related features, the study found that people who were self-employed had a higher risk for stroke and other cardiovascular diseases. Another study (Reis, Giroud) analyzed how environmental risk factors are associated with stroke [3]. The study accounted for all environmental factors such as different kinds of pollution, noise, and weather. The study concluded that adverse environmental factors increase the chances of an individual experiencing stroke. Apart from type of work and environmental aspects, an individual's personal detrimental habits such as smoking has been known to contribute to stroke. A systematic review by Pao and Jin inspected the correlation between smoking and stroke [1]. This article analyzed 14 studies involving a total of over 300,000 subjects. The study concluded that smoking has a dose dependent relationship with stroke. Moreover, the study concluded that smokers have a higher chance of experiencing stroke. Finally, a study that compared stroke patients who were and were not married and concluded that those who were unmarried had a risk of stroke [4].

In our study, we extend the current research and aim to explore the degree of correlation between lifestyle factors and stroke. To achieve this goal, we used two baseline machine learning algorithms – Logistic Regression and Decision Tree, along with a Neural Network. We used a stroke dataset that consisted of health related features as well as lifestyle features. Our aim is to first test the dataset on both baseline models while comparing and contrasting the results. We will then exclude the lifestyle features from the dataset and examine how that affects the accuracy of the model.

III. METHOD

a. DATA

We began exploring the data by comparing the medical features to the lifestyle features. The first thing we noticed was that the dataset was imbalanced where only 5% of patients had a stroke and the majority 95% did not. We addressed this skewness later when creating our models.

We began our exploration with the medical features, starting with age. The dataset included children, some even less than a year old. We found that the older the patient was, the higher the risk of stroke. With BMI, most of the patients with stroke had an average BMI of 30, where any level above 30 is considered to be obese. Though this is also the case with those without stroke, the average BMI of the entire dataset was 30. We have found that those with hypertension had a much higher risk of stroke than those without. This is also true for those with heart disease with a 3 to 4 times higher chance of risk. But no significance was found when comparing average glucose levels or gender.

Then we continued our exploration with the lifestyle features. We found that those who were self-employed had a higher risk compared to those who

worked for a private company or the government, supporting the current research on work type and risk of stroke. The patients that have ever been or were married had a higher risk of stroke than those who have never married. This finding challenges the current study on the effects of marital status on the risk of stroke where it was concluded that those who have never married had a higher risk of stroke. Another finding that challenges current research was the effects of smoking on stroke. We have found that those who formerly smoked had a higher risk of stroke than those that were currently smoking while the current research says the opposite. We have not found any significance with the type of residence which also challenged a study that found that urban dwellers had a higher risk of stroke than those who lived in rural areas.

b. PREPROCESSING

We made slight modifications to our data to produce a uniform set that would fit all three of our machine learning models. The first step was removing all rows that included "Not A Number" (NAN) values. Our models cannot perform on this data type and retaining these rows would introduce errors to our exploratory data analysis. Secondly, we converted many of our categorical columns to one-hot-notation including (gender, ever-married, work-type, residence-type, and smoking-status). Many machine learning models cannot perform on categorical data directly, therefore, we decided to transform these features into a numerical form. Third, we noticed our dataset had extremely skewed class proportions, the dataset only contained approximately 5% stroke compared to 95% no stroke. Initially, this caused some of our machine learning models to only produce a negative stroke output producing high accuracy with a high false negative rate. To combat this issue, we used a library function called random-oversample from imbalanced-learn to over-sample the stroke class bringing the final class ratio to 3:7. Our last step was removing the identification column from the dataset due to its lack of statistical relevance to the dataset.

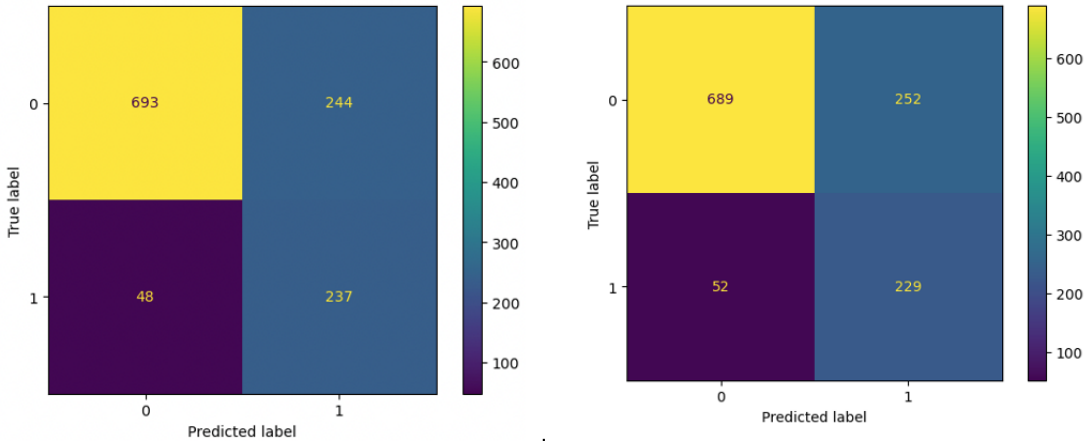
c. MODELS

We decided to compare two baseline Machine Learning (ML) models to a customized Neural Network. For the first two ML models we used logistic regression and a decision tree. We decided on these two models due to their vast differences. For example logistic regression assumes the dataset is linear separable and fits the data to a line whereas a decision tree does not and divides the data into regions. Our initial assumption was that the neural network would perform best due to its ability to handle data with and without linearities. For logistic regression we used the standard Scikit-learn with a max iteration of 1000 and the class-weight parameter set to 'balanced'. For the decision tree we used all

default settings from the Scikit learn library. Lastly, we designed a custom neural network using the pytorch library. We used a 21-node input later with an ReLu activation function moving to an 8-node hidden later with a sigmoid activation function ending with a one node output layer. We used several metrics to compare these models such as accuracy, AUC curve, AUC score, and a confusion matrix. Our second task was to compare models with and without lifestyle features. We divided the data into two sets one with all features and one with only medical features. Based on the exploratory data analysis we predicted the model with lifestyle features would perform better than the one without due to the small correlations each feature had with producing a stroke.

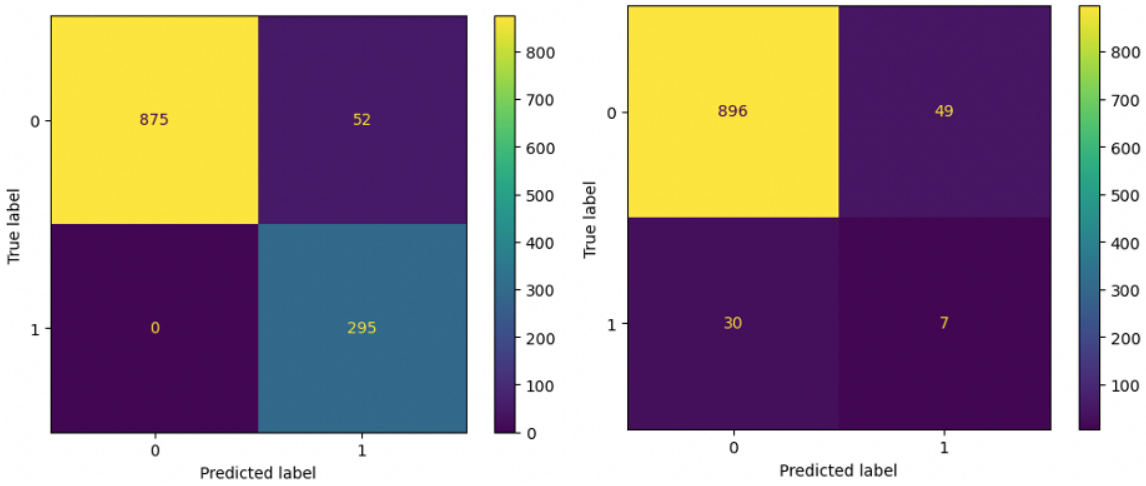
IV. RESULT

Figure 1. The image below shows the confusion matrix results for the logistic regression run on the original dataset (left) and the dataset excluding lifestyle features.



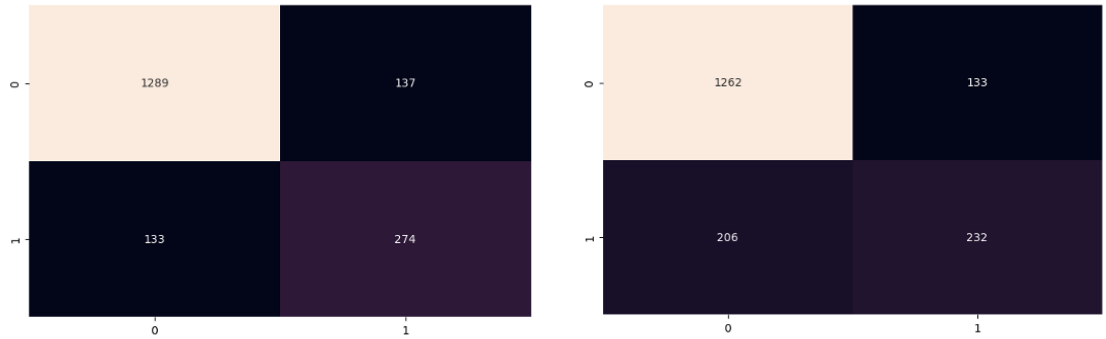
In our case, false negatives carried more weight as compared to other metrics. For the original dataset, the model resulted in a false negative rate of 16.8% compared to 18.5% for the modified dataset. Like mentioned above, false negatives pose more of a threat. A medical check up could declare an individual fine, when in reality the person's health could be in a precarious state. Hence, we can easily see that removing the lifestyle feature adversely impacted the accuracy of the model.

We noticed similar results with our Decision Tree. The image below shows the confusion matrix results for both the original dataset (left) and modified dataset (right).



As can be seen in the image above, our decision tree actually resulted in no false negatives for the original dataset, but had a false positive rate of around 81%. In this case as well, the accuracy of the model dropped after removing the lifestyle features.

Apart from the two baseline models - Logistic Regression and Decision Tree, we also tested our model on a Neural Network. Even though dropping the lifestyle features did not impact the neural network's accuracy in any way, there were still more false negatives observed for the dataset without lifestyle features. The network produced a false negative rate of 32.7% for the original set whereas the modified set had a false negative rate of about 47%. The confusion matrix results are shown below.



Original Dataset

Modified Dataset (No Lifestyle Features)

For both baseline models, the accuracy dropped when lifestyle features were removed. Whereas for the Neural Network, dropping lifestyle features didn't have any effect on the accuracy. The accuracy and AUC Scores for all three models are summarized below.

ML Model Comparison		
	Accuracy	AUC Score
Logistic Regression - All Features	0.76	0.85
Logistic Regression - No Lifestyle Features	0.73	0.84
Decision Tree - All Features	0.95	0.97
Decision Tree - No Lifestyle Features	0.91	0.56
Neural Network - All Features	0.81	NA
Neural Network - No Lifestyle Features	0.81	NA

V. CONCLUSION

In conclusion, stroke is a serious condition that affects close to a million Americans per year. Experiencing this devastating medical emergency can cause several side effects such as paralysis, speech and language problems, memory issues, and even visual impairment. While the majority of strokes are preventable, most studies only focus on medical markers rather than lifestyle choices. Our study focuses on alternative choices individuals can make related to stroke risk. From our analysis of the stroke dataset we have found a correlation of several lifestyle choices to increased risk of stroke. For example, self employed individuals are at a higher risk of stroke than individuals working for private companies or the government. We also found that individuals who were married had a higher risk of stroke than those never married. The most surprising result was that former smokers had a higher risk of stroke than those that currently smoke. Our next analysis was contrasting several machine learning models when comparing datasets with lifestyle features to a dataset without lifestyle features. Through this analysis we produced two sets of results: a comparison of three different machine learning models and the comparison of machine learning model accuracy when using datasets with and without lifestyle features. We found a decision tree model to be more accurate than a logistic regression model and a custom designed neural network. Lastly, we found datasets that include lifestyle features produce more accurate results when applied to machine learning models.

VI. REFERENCES

[1] B. Pan, X. Jin, L. Jun, S. Qiu, Q. Zheng, and M. Pan, “The relationship between smoking and stroke,” *Medicine*, vol. 98, no. 12, 2019.

[2] C. Krittanawong, A. Kumar, Z. Wang, U. Baber, and D. L. Bhatt, “Self-employment and cardiovascular risk in the US general population,” *International Journal of Cardiology Hypertension*, vol. 6, p. 100035, 2020.

[3] J. Reis, M. Giroud, and Y. Kokubo, “Environmental risk factors for stroke and cardiovascular disease,” *Encyclopedia of Cardiovascular Research and Medicine*, pp. 238–247, 2018.

[4] Q. Liu, X. Wang, Y. Wang, C. Wang, X. Zhao, L. Liu, Z. Li, X. Meng, L. Guo, and Y. Wang, “Association between marriage and outcomes in patients with acute ischemic stroke,” *Journal of Neurology*, vol. 265, no. 4, pp. 942–948, 2018.

[5] “Brain Basics: Preventing Stroke”.*National Institute Of Health*. https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-preventing-stroke

[6] “How many people are affected by/at risk for stroke?”. *National Institute Of Health*, 2022.