Project Milestone-1

DSC-680

Debasish Panda

Introduction:

A stroke, or brain attack, happens when blood flow to our brain is stopped. It is an emergency situation. The brain needs a constant supply of oxygen and nutrients in order to work well. If blood supply is stopped even for a short time, this can cause problems. Brain cells begin to die after just a few minutes without blood or oxygen. When brain cells die, brain function is lost. We may not be able to do things that are controlled by that part of the brain. For example, a stroke may affect our ability to move, speak, eat, think and remember, controlling emotions, controlling bowel and bladder etc. In other words, a stroke can happen to anyone at any time.

- According to CDC, every year more than 795,000 people in United States have a stroke. Every 4 minutes someone dies of stroke.
- Even though stroke is a leading cause of death for Americans, but the risk of having a first stroke is nearly twice as high for blacks as for whites, and blacks have the highest rate of death due to stroke.
- Recurrent stroke is frequent; about 25 percent of people who recover from their first stroke will have another stroke within 5 years.

Since stroke is something that can happen to anyone at any age, if we can use some of the prediction modelling techniques to predict whether a given person is likely to have a stroke or not based on the risk factors, we discussed then we can save a lot more lives.

Scope:

The main objective of this paper is to design a robust system that works efficiently and will be able to predict the possibility of Stroke accurately. This paper uses the dataset available in Kaggle with 5110 observations.

Link to dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Here are the attributes of the dataset explained:

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) stroke: 1 if the patient had a stroke or 0 if notBelow is the link to the actual data set.

Types of models I plan to use:

I'm planning to use Logistic regression on this data set as it is the appropriate regression analysis to conduct when the dependent variable is binary. Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

Plan to evaluate the results:

I plan to follow the below steps to evaluate the results.

1. Load the data set into a data frame.
2. Perform the EDA to understand the characteristics of the data set.
3. Evaluate the correlation between the variables in the dataset.
4. Divide the data set into a train and test data set and apply the logistic regression model.
5. Create a confusion matrix to show the performance of the model to evaluate the predicted values from the model vs. the actual values from the test dataset.

I hope to learn:

By using this model, I plan to learn the nuances of the logistic regression and evaluate my strengths and weaknesses while working on a model. I expect to curate an effective, accurate and working prediction model which can predict the possibility of occurrence of a stroke.

Ethical Considerations:

Since we are to deal with the medical records of patients, we may need to provide more attention towards any Personally Identifiable Information is that could identify an individual is not misused. This could be

the names, home address, date of birth, email address etc. One more item that we should be taking into account is not to be biased by gender and race.

Risks with the proposals:

Logistic regression models are not perfect and have their disadvantages, "Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So, on high dimensional datasets, Regularization techniques should be considered to avoid over-fitting.

At this point of time, I am guessing that there are enough number of observations and the contents are accurate. However, if we encounter a situation where there are not enough data points to create a predictive model then it would be a disaster. Hence, I have come up with a contingency plan of customer churning for telecom industry.

References:

1. NHLBI. (March24, 2022). What is a Stroke? National Heart, Lung and Bone Institute. Retrieved December 12, 2022, from https://www.nhlbi.nih.gov/health/stroke
2. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset