Project 1 White Paper

Kristie Kooken

DSC 680-T301

Due: April 7, 2024

Business Problem:

The annual estimated costs of loss of productivity from the chronic diseases of diabetes and hypertension is staggering, about 600 billion US dollars (www.cdc.gov). This investigation will explore how different health behaviors and physical characteristics (e.g., age and sex) as well as diabetes and hypertension diagnoses can predict who can be at risk for stroke.

Background/History:

Stroke is a leading cause of death and severe long-term disability (www.heart.org). Diabetes is a well-established risk factor for stroke (Chen, et al., 2017), and for those with diabetes, the risk of stroke is two times higher than those who do not have diabetes (www.diabetes.org). Likewise, high blood pressure can often have no symptoms and if untreated, may lead to conditions such as heart disease and stroke (www.mayoclinic.org). Hypertension plays a part in about 50% of all strokes (www.stroke.org.uk). Having a better understanding of risk factors associated with stroke can help further develop interventions to help those at risk for this serious life-threatening medical condition.

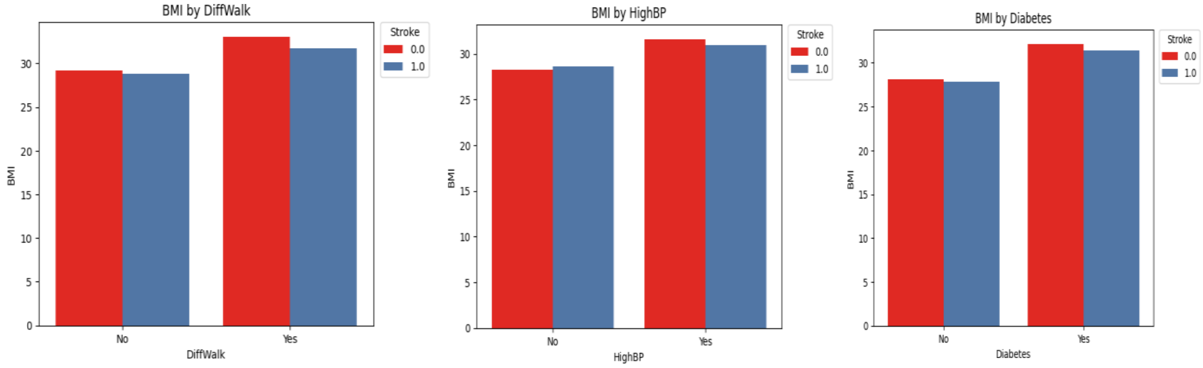
Data Explanation:

Using a dataset from kaggle.com that is based on the 70,692 survey responses from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 (www.kaggle.com). This survey collects health behavioral risk factor data and is an established data collection tool since 1984 (www.cdc.com). All data cleaning and analyses were conducted in python using Jupyter Notebook. This dataset consists of 18 columns and the data dictionary for these columns is in Appendix 1.

Method:

This data was scrubbed by ensuring there were no missing values and removing any duplicates. Once this was completed, exploratory data analysis (EDA) was conducted to check the distribution of each variable as well as determine if there were any outliers. Variable distributions were as expected, and no outliers were found. The resulting dataset had 64,020 rows. The bar charts for BMI by difficulty in walking, high BP or diabetes each show how stroke is higher in each of these categories. Otherwise, the EDA did not show other variables to have much difference when examined by stroke status.

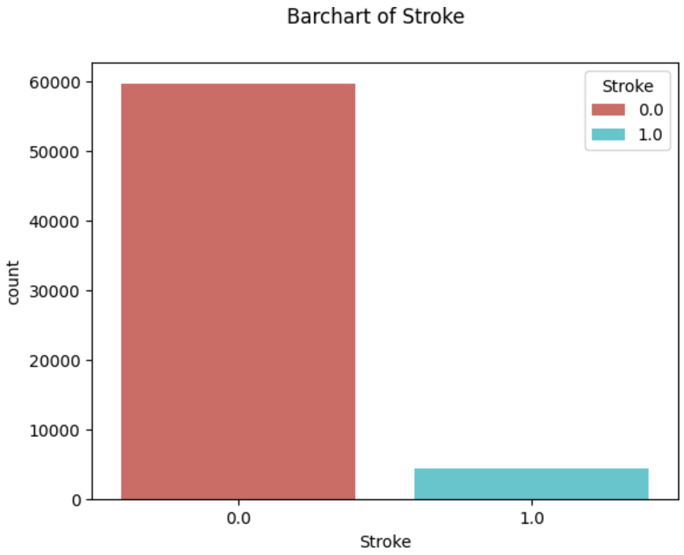
Figure 1.



Analysis:

To prepare the data for analysis, those categorical variables with more than 2 levels were dummied to convert each category into a binary numerical variable. The next step was to use SMOTE to up sample the number of strokes in the dataset. SMOTE, Synthetic Minority Oversampling Technique, was employed on the train data to create a balance dataset on the variable Stroke. SMOTE allowed for two benefits, 1) to create synthetic samples from the minor class (Yes - Stroke) instead of creating copies and 2) randomly choosing one of the k-nearest-neighbors and using it to create similar, but different, new observations (www.towardsdatascience.com).

Figure 2.



The next step was to investigate if columns could be dropped from the analysis using recursive feature elimination (RFE). Of the twenty-eight columns in total, 8 were dropped using this method. Columns for high cholesterol, cholesterol check in the past 5 years, smoker, heart disease history, heavy alcohol consumption, difficulty in walking, high BP, diabetes and age remained. A logistic regression model was run on this final dataset with stroke as the dependent variable. Confusion matrix and ROC plots were created.

Conclusion

Accuracy for any health model is extremely critical. The logistic regression produced a model with 80% accuracy, which is an acceptable threshold. Likewise, the precision and recall values for this model are also acceptable (See Table 1).

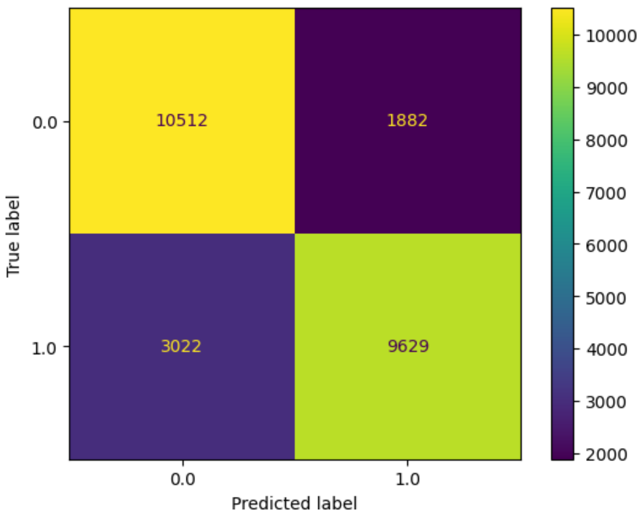
Table 1.

Classification Table

Predictor value	Precision	Recall	F1-score	Support
0	0.78	0.85	0.81	12394
1	0.84	0.76	0.80	12651

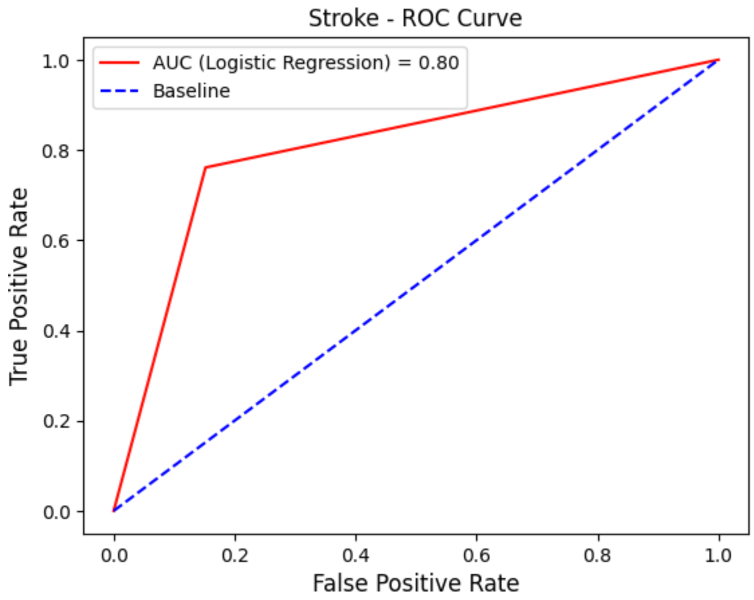
The confusion matrix is a visualization of the true negatives, false positives, false negatives, and true positives which are the correct and incorrect number of predictions for each class. This matrix shows 10512 of true negatives, 1882 of false positives, 3022 of false negatives and 9629 of true positives.

Figure 3.



The ROC curve is used to show how the logistic regression model discriminates between the positive and negative classes at different classification thresholds. The ROC plot shows a line above the 0.5 threshold, with the AUC score of 0.80 indicates an acceptable ability to discriminate between the positive and negative classes.

Figure 4.



Assumptions

Data source assumptions are that the data presented here is representative of the source survey and all data are correctly collected and scrubbed. Likewise, there is an assumption that using one year of data is sufficient for model development.

Analysis assumptions made are that using SMOTE is the best way to handle imbalance in the dataset and that this did not cause any overfitting of the model.

Limitations

Limitations of this analysis include using only one year of data from this national survey, not knowing the full extent of cleaning that was done on the data prior to being posted on kaggle.com.

Analysis limitations include using one type of modeling technique to predict stroke. Likewise, it would have been beneficial to be able to include a time or duration of the conditions of stroke, diabetes and hypertension in order to develop a survival model. Since diabetes and hypertension can lead to stroke, it is of benefit to understand at what stage in these chronic conditions could the possibility of stroke occur and how different interventions can alter that outcome.

Challenges

A challenge for me was that I initially ran a different set of analyses which did not yield a good model. This was not expected as I assumed the analysis would perform to an acceptable level. Thus, I had to reassess my data and any shortcomings (imbalance was not expected necessarily) and create a new plan.

Future Uses/Additional Applications

It would be great to be able to further develop this work and this type of model and create a larger application. There is a great need to find workable interventions for chronic health conditions like diabetes and hypertension. Also, it would be a great benefit for all to use prediction to ward off serious medical conditions like stroke.

Recommendations

Recommendations include getting multiple years of survey data to build a more robust model and to have a deeper investigation of the performance of different independent variables included in this model.

Implementation Plan

Currently, there is no implementation plan as this research effort is preliminary and additional data and modeling is needed to develop a robust intervention to aid in stroke prevention.

Ethical Assessment

It is important to ensure that the survey data represents all people who suffer from these chronic diseases and not only those who are willing to participate in this research. Likewise, results of this survey may not be representative of all populations who have these chronic diseases like diabetes and hypertension, there are limited demographics collected so it will be a challenge to know how balanced the sample is. Also, it is important to keep in mind that this data is limited to survey data which is self-reported from the participants and could differ from what they do in everyday life. Lastly, as with most results, it is important to ensure we have the full picture. What is presented here is a single year’s worth of data thus it is challenging to draw conclusions and important to keep in mind a lot more work is needed in order to have a production level model to develop any recommendations or interventions for stroke.

References:

Behavioral Risk Factor Surveillance System (accessed 2024, March 17). The Center for Disease Control and Prevention. https://www.cdc.gov/brfss/about/brfss_faq.htm.

Chen R, Ovbiagele B, Feng W (2016). Diabetes and Stroke: Epidemiology, Pathophysiology, Pharmaceuticals and Outcomes. Am J Med Sci. 2016 Apr;351(4):380-6. doi: 10.1016/j.amjms.2016.01.011. PMID: 27079344; PMCID: PMC5298897.

Diabetes, Hypertension and Stroke Prediction (accessed 2024, March 17). Kaggle. https://www.kaggle.com/datasets/prosperchuks/health-dataset/data.

Get serious about stroke prevention (accessed 2024, March 31). American Diabetes Association, Diabetes Complications. https://diabetes.org/about-diabetes/complications/stroke.

Health Topics – Diabetes (accessed 2024, March 17). The Center for Disease Control and Prevention. https://www.cdc.gov/policy/polaris/healthtopics/diabetes/index.html#:~:text=The%20total%20estimated%20cost%20of,90%20billion%20in%20reduced%20productivity.

Health Topics – High Blood Pressure (accessed 2024, March 17). The Center for Disease Control and Prevention. https://www.cdc.gov/policy/polaris/healthtopics/highbloodpressure/index.html#:~:text=Economic%20Burden,to%20%24198%20billion%20each%20year.

High blood pressure (accessed 2024, March 31). Stroke Association. https://www.stroke.org.uk/stroke/managing-risk/high-blood-pressure#:~:text=High%20blood%20pressure%20plays%20a,your%20risk%20of%20a%20stroke.

High blood pressure (hypertension) (accessed 2024, March 17). The Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410?utm_source=Google&utm_medium=abstract&utm_content=Hypertension&utm_campaign=Knowledge-panel.

Li, S (2017). Building A Logistic Regression in Python, Step by Step. Towards Data Science. https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8.

Appendix 1:

Table of variables with their respective decodes used in analysis.

Variable	Decode
Age	13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
Sex	Patient's gender (1: male; 0: female).
HighChol	0 = no high cholesterol 1 = high cholesterol
CholCheck	0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
BMI	Body Mass index
Smoker	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
HeartDiseaseorAttack	Coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
PhysActivity	Physical activity in past 30 days - not including job 0 = no 1 = yes
Fruits	Consume Fruit 1 or more times per day 0 = no 1 = yes
Veggies	Consume Vegetables 1 or more times per day 0 = no 1 = yes
HvyAlcoholConsump	Adult men >=14 drinks per week and adult women>=7 drinks per week: 0 = no 1 = yes
GenHlth	Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
MentHlth	Days of poor mental health scale 1-30 days
PhysHlth	Physical illness or injury days in past 30 days scale 1-30
DiffWalk	Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
Stroke	Have you ever had a stroke. 0 = no, 1 = yes
HighBP	0 = no high, BP 1 = high BP
Diabetes	0 = no diabetes, 1 = diabetes

Appendix 2:

Ten questions an audience would ask:

1) Did you have to do many data transformations?
2) What other types of analyses would you consider exploring with these data?
3) Have you considered expanding your variable selection from the Behavioral Risk Factor Surveillance System and addressing gaps in your analysis?
4) For your analysis, you mentioned initially going a different route, why did you change course?
5) Do you think your model can truly predict who will have a stroke?
6) How dirty was your data?
7) Did you consider regrouping certain variables like age and making fewer categories?
8) From your ethical considerations, can you explain more about sampling techniques for these surveys?
9) What other demographics do you think could be beneficial?
10) Did you consider creating a surrogate for time based on different research information related to diabetes and hypertension?