UNIVERSITY OF THE FRASER VALLEY

“BRAIN STROKE PREDICTION”

COMP 431
DATA MINING
DR. DAVID CHU

VINODKUMAR MANDORA
300179952

INTRODUCTION

Technology has helped us witness some amazing advancements in the field of medicine. With the collection of multiple datasets of medical records, data mining techniques have largely helped identify trends in the datasets. The analysis has helped the medical practitioners to make an accurate prognosis of medical conditions which led to improved healthcare conditions and reduced treatment costs.

In 2020, 1 in 6 deaths from cardiovascular disease was due to stroke. Every 40 seconds, someone in the United States has a stroke. Every 3.5 minutes, someone dies of stroke. Every year, more than 795,000 people in the United States have a stroke. About 610,000 of these are first or new strokes. About 185,000 strokes nearly 1 in 4 are in people who have had a previous stroke. About 87% of all strokes are ischemic strokes, in which blood flow to the brain is blocked. Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018. This total includes the cost of health care services, medicines to treat stroke, and missed days of work.

Assisting the medical practitioners to identify the onset of disease at an earlier stage is a huge help when it comes to saving lives. For this project we mainly focus on strokes and to identify some important factors associated to its occurrence. Its really important identify the key factors that lead to a stroke prediction. We aim to provide a systematic analysis of the various patient records for the purpose of stroke prediction.

OBJECTIVE

A model to predict an early onset of a stroke is required to minimize life loss as well as to reduce cost of healthcare. So, we use the available data and analyze it to create a model to predict if a patient will have a brain stroke using SAS

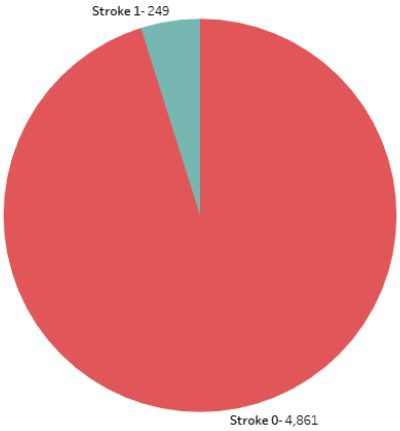
DATA

The dataset has been collected from Kaggle (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) uploaded by Fede Soriano and consists of 12 variables. The dataset consists of 5110 records. Id variable is the primary key and would not be used in the creation of the model. Stroke variable is our target variable which consists of binary values 0 and 1. 1 means the patient suffered a stroke and 0 indicates they didn't. Out of the remaining 10 variables we have 5 binary variables, 3 quantitative variables and 2 categorical variables.

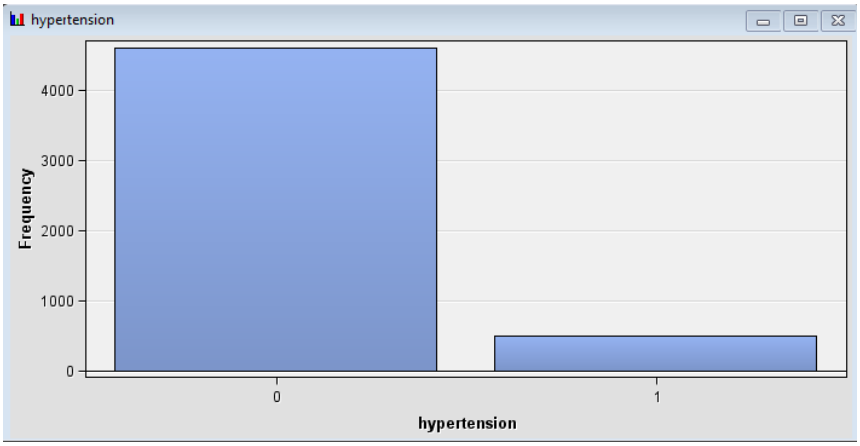
S.No.	ATTRIBUTE	DEFINITION
1	ID	Unique patient ID
2	GENDER	Gender of the Patient [Male, Female]
3	AGE	Age of the patient in years
4	HYPERTENSION	1- For patients suffering from hypertension 0- If not
5	HEART_DISEASE	1- For patients suffering from a heart disease 0- If not
6	EVER_MARRIED	Yes - If patient is married No - If not
7	WORK_TYPE	[Government job, Never Worked, Private, Children, Self-employed]
8	RESIDENCE_TYPE	Type of residence [Urban, Rural]
9	AVG_GLUCOSE_LEVEL	Average glucose level of the patient
10	BMI	BMI of the patient
11	SMOKING STATUS	Date of the month of the 2018 year
12	STROKE	1 -If the patient suffered a Stroke 0 - If not

DATA EXPLORATION

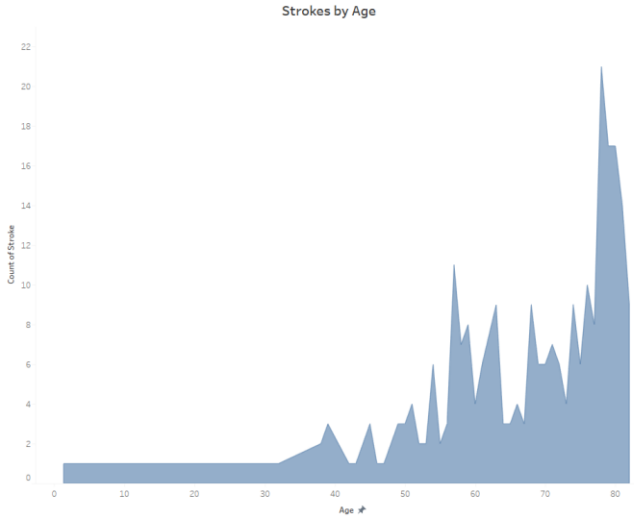
In the pie chart below we can see that only 4.8%(249) of the total patients suffered a stroke.



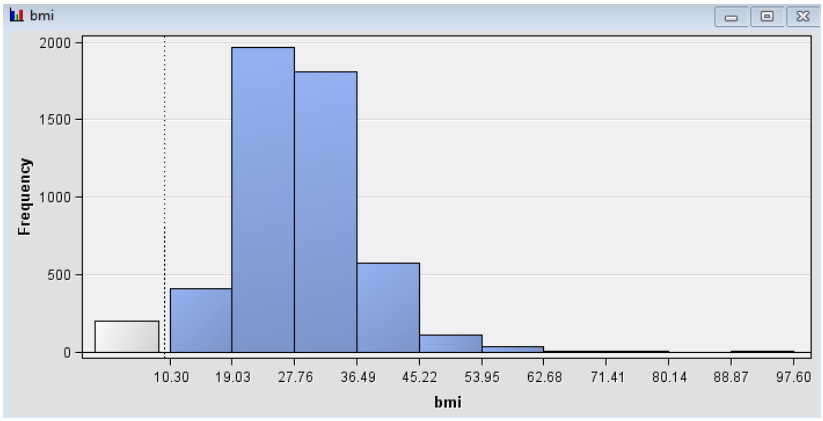
The bar chart below shows that 498 out of 5110 patients suffer from hypertension



We can also notice in the visualization below that more people start having a stroke after the age of 55

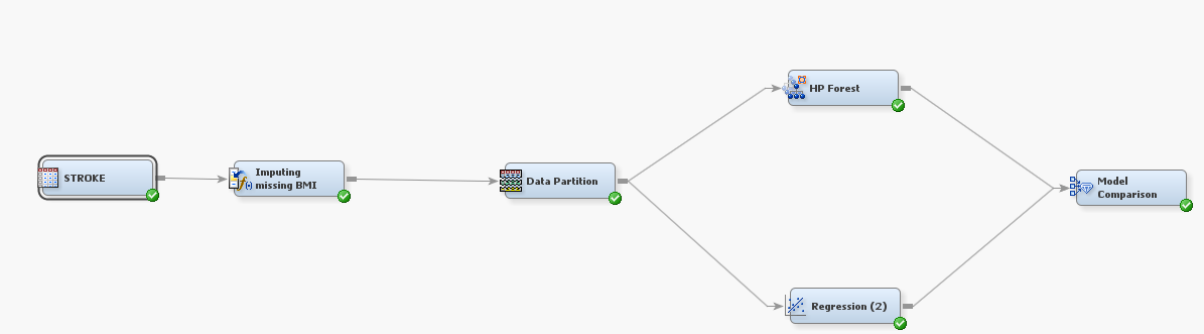


DATA CLEANING



In the frequency chart for the BMI variable we can notice that there are about 201 missing values. Removing these rows can lead to a loss of useful information so, we impute the mean BMI value 28.8932 for the missing rows.

MODEL CREATION



Once the dataset is cleaned we make a 50-50 partition of the dataset into training and validation data.the training data is used to form a model and the validation data is used to check how well the model performs on unknown data

RANDOM FOREST CLASSIFIER

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement. We use this to create an appropriate model. The random forest classifier gives us the following loss reduction variable importance table which shows us a list of variables ordered by their importance

We can see that age, heart_disease, hypertension and avg_glucose level have a significant relationship with our target variable stroke

Loss Reduction Variable Importance

Variable	Number of Rules	Gini	OOB Gini	Valid Gini	Margin	OOB Margin	Valid Margin
age	182	0.005329	0.00327	0.00444	0.010658	0.00835	0.00946
heart_disease	83	0.001724	0.00112	-0.00003	0.003447	0.00276	0.00149
hypertension	58	0.000543	0.00032	0.00023	0.001085	0.00074	0.00089
avg_glucose_level	37	0.000856	0.00028	-0.00020	0.001712	0.00112	0.00120
ever_married	33	0.000174	0.00004	0.00007	0.000348	0.00021	0.00024
work_type	9	0.000039	-0.00001	0.00001	0.000078	-0.00001	0.00002
gender	1	0.000005	-0.00001	-0.00001	0.000009	-0.00002	-0.00001
IMP_bmi	17	0.000270	-0.00007	-0.00038	0.000539	0.00017	-0.00011
smoking_status	8	0.000052	-0.00008	-0.00003	0.000104	-0.00006	0.00002
Residence_type	8	0.000041	-0.00009	-0.00006	0.000083	-0.00010	-0.00007

As shown in the table below the random forest classifier model does a pretty job on the validation data with a misclassification rate of 0.048551

Event Classification Table

Data Role=TRAIN Target=stroke Target Label=' '

False Negative	True Negative	False Positive	True Positive
124	2430	0	0

Data Role=VALIDATE Target=stroke Target Label=' '

False Negative	True Negative	False Positive	True Positive
125	2431	0	0

LOGISTIC REGRESSION

Logistic regression is a type of regression where a logit transformation is applied on the odds. That is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi))

As our target variable is binary a logistic regression model can be a really good idea to reach a solution model. We will be following a stepwise approach towards model building.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-6.8210	0.5430	157.77	<.0001		0.001
age		1	0.0629	0.00709	78.61	<.0001	0.7803	1.065
avg_glucose_level		1	0.00551	0.00162	11.56	0.0007	0.1391	1.006
heart_disease	0	1	-0.3142	0.1271	6.11	0.0134		0.730
hypertension	0	1	-0.2348	0.1141	4.24	0.0396		0.791

In the table above we see the final model that we form after applying a logistic regression. The variables age avg_glucose level, heart_disease and hypertension were again found to be of significant importance to predict a stroke. The significance of these variables can be tested by the likely hood ratio test where,

H0 : B.hat=0

H1 : B.hat !=0

In the probability column shown in the table we can see that all the probability values are less than 0.5. So we have sufficient evidence to reject the Ho for all variables and we can say we are 95% confident that the variables age, avg_glucose_level, heart_disease and hypertension have a significant effect on stroke variable. The final model can be written as below

$$Probability\ of\ stroke = \frac{1}{1 + e^{-(0.001 + 1.065(age) + 1.006(AvgGlucoseLevel) + 0.730(HeartDisease_0) + 0.791(Hypertension_0))}}$$

The below picture shows us how good of a job the logistic model does on the dataset. It achieves a misclassification rate of 0.048551 which is a pretty good result.

Event Classification Table

Data Role=TRAIN Target=stroke Target Label=' '

False Negative	True Negative	False Positive	True Positive
122	2428	2	2

Data Role=VALIDATE Target=stroke Target Label=' '

False Negative	True Negative	False Positive	True Positive
125	2431	.	.

MODEL COMPARISION

So we come up with two really good models with relatively different approaches. The misclassification rate is surprisingly similar for both the models however the logistic regression model edges out ahead with a better Average squared error rate which is why we choose the logistic regression model as our final model.

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg2	Regression (2)	0.048905	0.041941	0.048551	0.042663
	HPDMForest	HP Forest	0.048905	0.041147	0.048551	0.043170

INTERPRETATION [LOGISTIC MODEL]

To interpret the logistic regression model we can say that keeping other factors constant:

- A 1 unit increase in Age multiplies the chances of a stroke by 1.065 times
- A 1 unit increase in average glucose level multiplies the chances of a stroke by 1.006 times
- Patients without heart disease have 0.73 times the odds of the person with a heart disease of having a stroke
- Patients without hypertension have 0.791 times the odds of the person with a heart disease of having a stroke

CONCLUSION

To conclude we can say that we came up with two really good models with the logistic regression model performing slightly better than the random forest classifier. The variables age, avg_glucose_level, heart_disease and hypertension were found to be of importance in both the models. However, we also notice that this models could be further strengthened by having a larger dataset as the more the data the more knowledgeable the model is. We can also add more variables to the dataset like history of a predecessor suffering from strokes etc.