Stroke Prediction

An Exercise in Machine Learning and Stroke Probability Predictions

Girish E
Raj Jain

Executive Summary

Cerebrovascular accidents (strokes) in 2020 were the 5th leading cause of death in the United States.

The objective of this activity was to develop a model which can reliably predict the likelihood of a stroke using patient input information.

Hypothesis

A reliable predictive model can be developed if the data and stroke key attributes are correctly identified and prepared for the machine learning process. The importance of features generated by the model selected will be compared against the stroke risk factors identified by the American Stroke Association. If the attributes are correctly identified by the model, the hypothesis will be considered validated.

Data Selection

A dataset from Kaggle was selected for the machine learning process. The data was reviewed to identify trends and cleanup requirements. The primary data cleanup activities identified were to address "N/A" values associated with body mass index and "Unknown" smoker status. The "N/A" values represented 3.9% of the dataset and were addressed by using the mean body mass index and assigning that to the "N/A" values. The "Unknown" smoker status represented 30.4% of the dataset. Literature review verified that "Unknown" values were considered an accepted data point and therefore the "Unknown" values were left as presented in the raw data.

Data Preparation

Review of the data identified most of the Yes/No type answers for personal health questions, including if the person had a stroke, were heavily biased to the "No" side. To ensure an effective model learning process, Synthetic Minority Oversampling Technique (SMOTE) was used to synthetically balance the Yes/No results for stroke. SMOTE selects samples in the minority class that are close and then draws lines between them. New sample points are located on these lines. Other data preparation steps included One-Hot Encoding.

Model Selection and Machine Learning

Linear and Tree models were evaluated. The criteria used to select the model was one that did not overfit the data and had the best overall performance for f1-scores and recall for the likelihood of a stroke. After the evaluation process was completed, a Linear Model, LogisticRegression was selected. The results were saved and tested against live data after the model selection and machine learning process was completed.

Hypothesis Validation

The live testing of data verified that a reliable predictive model could be created if the data and stroke key attributes are correctly identified and prepared for the machine learning process.

Table of Contents

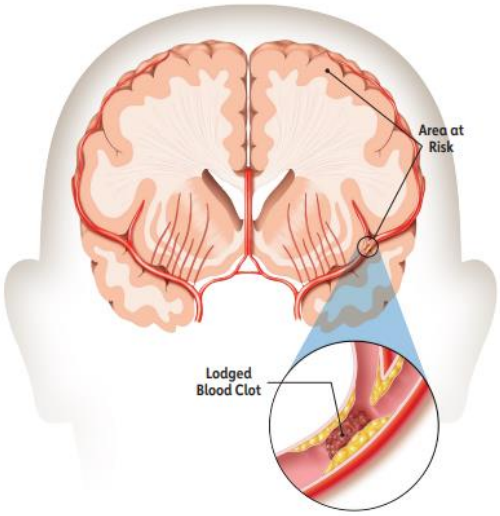
Introduction

The objective of this activity is to develop a preliminary screening tool which can be used to identify the likelihood of an individual having a stroke based on general contributing attributes. Data [1] from Kaggle was used as the basis for the predictive model.

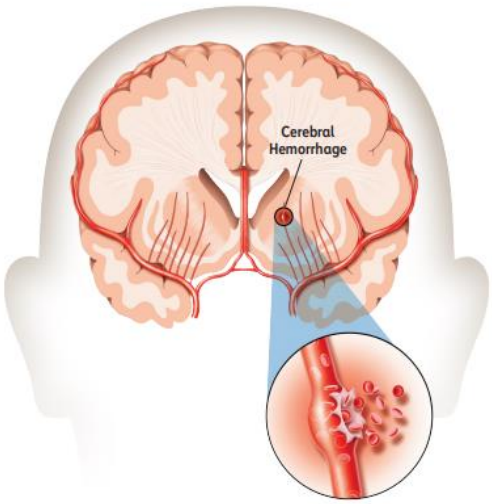
Cerebrovascular accidents (strokes) in 2020 were the 5th leading cause of death [2] in the United States.

A stroke occurs when the blood supply to a region of the brain is suddenly blocked or when a rupture occurs starving the brain cells of oxygen and nutrients. Blockage obstructing the flow of blood to a region of the brain is called an ischemic stroke and accounts for 87% [3] of all strokes. The rupturing of a blood vessel is called a hemorrhagic stroke and accounts for 13% [4] of all strokes.

Ischemic Stroke



Hemorrhagic Stroke



Source of Images [5]

The dataset used for the predictive model did not identify the type of stroke for each respective individual. To stay consistent with the dataset, the general word stroke will be used to describe the occurrence being predicted. A third category of stroke called a transient ischemic attack (TIA), or "mini stroke", caused by a temporary clot can also occur. The TIA has contributing factors similar to those of the ischemic and hemorrhagic stroke and is included in the general term stroke when identifying a potential outcome.

Per the American Stroke Association, 80% of strokes are preventable [6].

Hypothesis

By using data associated with stroke victims, a predictive model will be developed to identify the likelihood of a stroke.

Hypothesis: A reliable predictive model can be developed if the data and stroke key attributes are correctly identified and prepared for the machine learning process. The importance of features generated by the model selected will be compared against the stroke risk factors identified by the American Stroke Association. If the attributes are correctly identified by the model, the hypothesis will be considered validated.

Basis Risk Factors from American Stroke Association common to the dataset.

- High Blood Pressure
- Smoking
- Diabetes
- Obesity
- Heart Disease
- Age (cannot be controlled)
- Gender (cannot be controlled)

Data Source

The attributes with the dataset are:

- id: a unique identifier for each set of information
- gender: “Male, “Female”, “Other”
- age: age of the patient
- hypertension: 0 assigned if hypertension not present, 1 if patient has hypertension
- heart_disease: 0 assigned if heart disease not present, 1 if patient has heart disease
- ever_married: “No” or “Yes”
- work_type: “children”, “Govt_job”, “Never_worked”, “Private”, or “Self-employed”
- Residence_type: “Rural” or “Urban”
- avg_glucose_level: average glucose level in blood
- bmi: body mass index
- smoking_status: “formerly smoked”, “never smoked”, “smokes”, or “Unknown”
- stroke: 0 if patient has not had a stroke, 1 if patient has had a stroke

Data Review

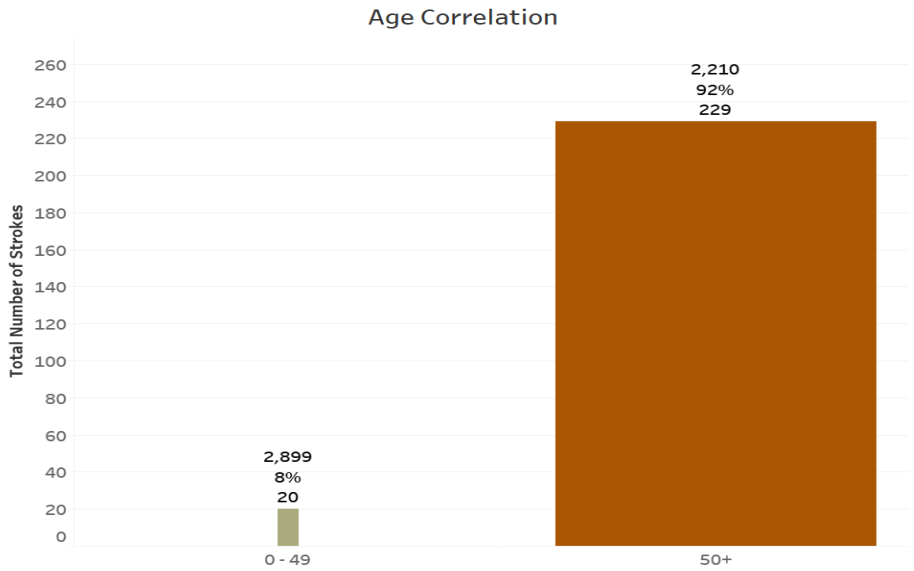
The raw dataset for machine learning consists of 5110 unique rows. Each row contains patient information designated by a unique id.

There were 2,994 (58.60%) “Females”, 2,115 (41.40%) “Males” and 1 “Other” in the gender attribute. The “Other” gender was dropped from the dataset for a resulting dataset of 5,109 unique rows.

Data Review				
Data Attribute	Female		Male	
	Count	Percent of gender	Count	Percent of gender
Had a stroke (Y)	141	4.7 %	108	5.1 %
Considered diabetic risk	230	7.7 %	204	9.6 %
Have heart disease (Y)	113	3.8 %	163	7.7 %
Have hypertension (Y)	276	9.2 %	222	10.5 %
Considered obese	1,115	37.2 %	805	38.1 %
Married (Y)	2,001	66.8 %	1,352	63.9 %
Live in Urban areas (Y)	1,529	51.1 %	1,067	50.4 %
Never smoked	1,229	41.0 %	663	31.3 %
Formerly smoked	477	15.9 %	407	19.2 %
Currently smoke	452	15.1 %	337	15.9 %
Unknown smoking status	836	27.9 %	708	33.5 %
Age: 0-19	480	16.0 %	486	22.9 %
Age: 20-39	791	26.4 %	412	19.4 %
Age: 40-49	450	15.0 %	280	13.2 %
Age: 50-59	472	15.7 %	362	17.1 %
Age: 60-69	352	11.8 %	269	12.7 %
Age: 70-79	336	11.2 %	233	11.0 %
Age: 80+	113	3.8 %	73	3.4 %

Trends identified in the dataset:

- 92% of strokes occur over the age of 50.



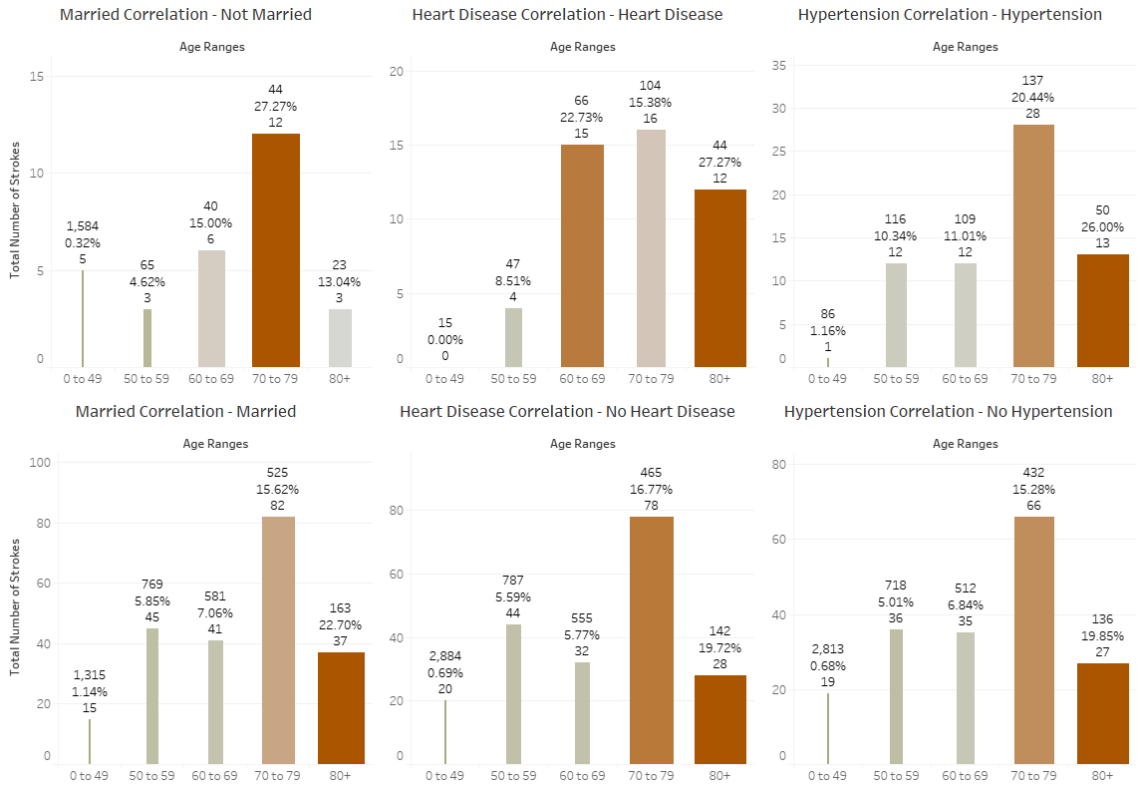
- The charts below are used to represent the general data characteristics defined by Yes/No answers. The attributes are in stacked panes for comparison purposes and the wider and darker brown bars indicate the higher normalized values.

  Comorbidities Heart Disease and Hypertension, along with falling under Not Married generally have a higher percentage of strokes as age increases when the data is normalized for the specific age ranges.

  The values associated with each bar are number of individuals in the designated age range, the normalized percentage of strokes in that range and the total number of individuals suffering a stroke in that range.

  The lower Percentage of Not Married for Age Range 80+ could include individuals whose spouses have died and therefore where not married at the time of the stroke. The phrasing of the question and related answer could have inadvertently redirected the overall results for the age range.

  Age Range 80+ generally has the highest percentage of strokes on a normalized basis.



- Comorbidities BMI, Glucose (Blood Sugar) and Smoker Status data are represented in the chart below. BMI (Overweight and Obese), Glucose (Diabetic Risk and Diabetic) and Smoker Status (Formerly Smoked and Current Smoker) have the highest normalized stroke percentages. The wider and darker brown bars indicate the higher normalized values.

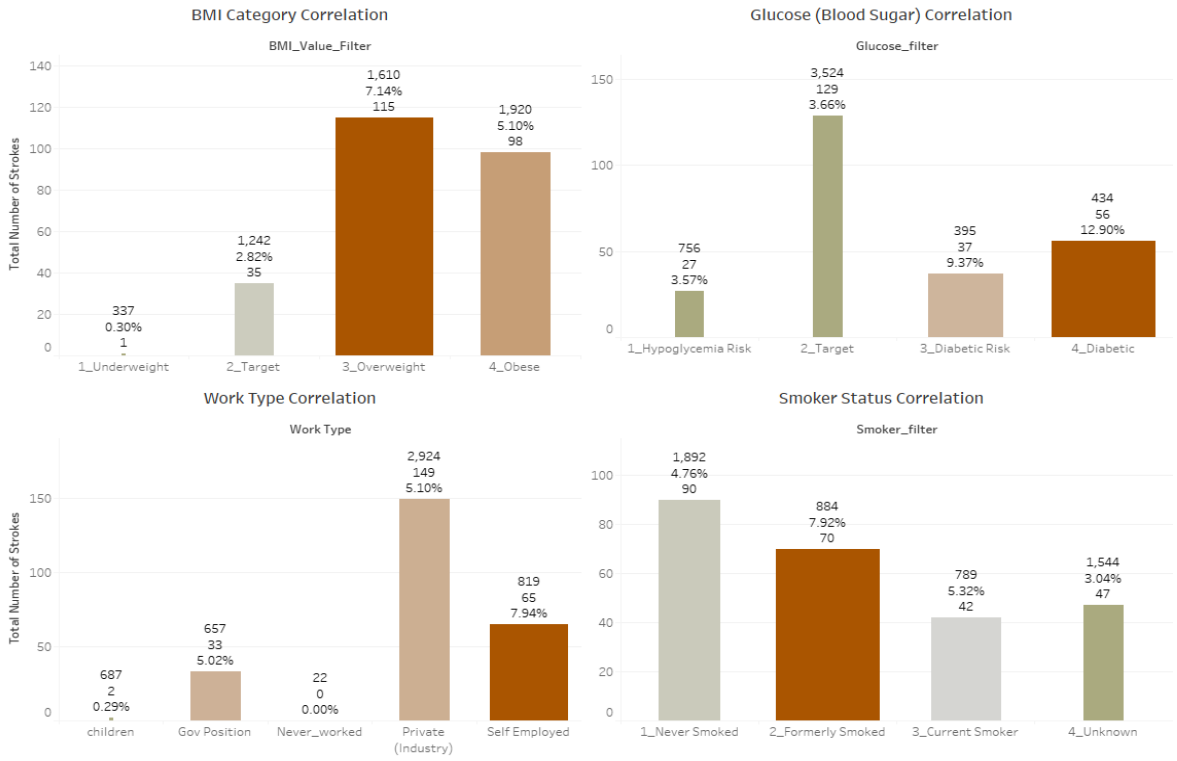
  The BMI Categories for data visualization are:

  - Underweight: BMI < 18.5
  - Target: 18.5 <= BMI < 25
  - Overweight: 25 <= BMI < 30
  - Obese: BMI > 30

  The glucose data presented in the dataset is average glucose value. Depending on when a person has eaten, a glucose value can have significant swings in values. Therefore, to create a Glucose Category that can used to filter the data, a blending of Fasting, Just Eaten and Several Hours after eating ranges were merged into ranges.

  The Glucose Categories for data visualization are:

  - Hypoglycemic: average glucose <= 70
  - Target: 70 < average glucose < 140
  - Diabetic Risk: 140 <= average glucose <= 200
  - Diabetic: > 200

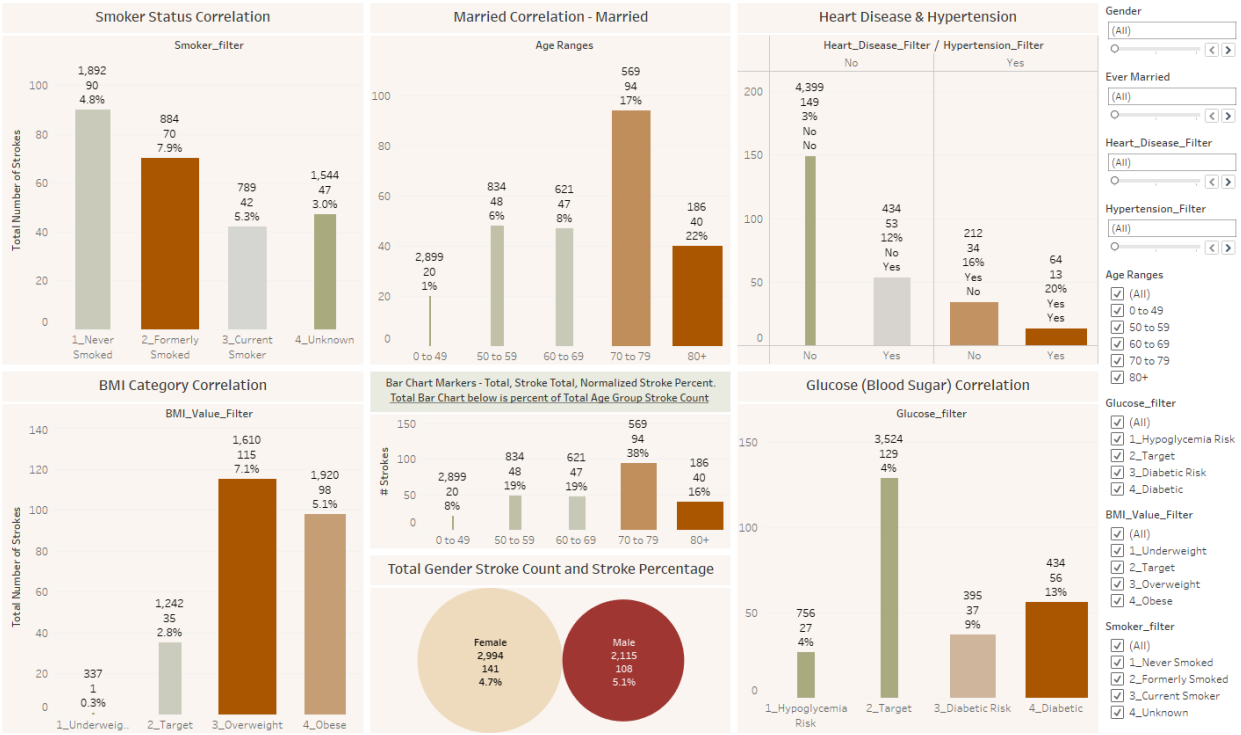


Visualizations

The data is represented in six different panes and one meet the data pane.

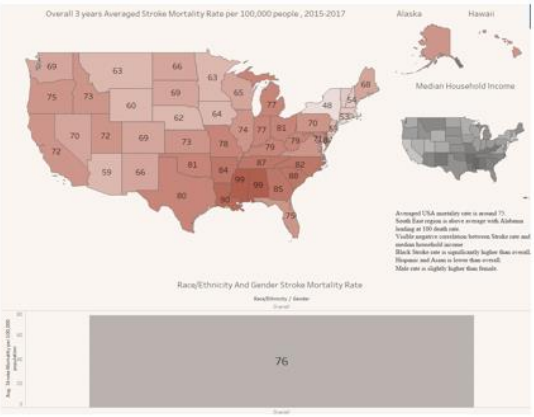
The main comorbidities are presented in five of the panes. The markers on each bar represents the total count, stroke total and stroke normalized percentage for the respective filter settings. The sixth pane lists gender, total count of the respective gender, total strokes, and stroke normalized percentage for the filter settings. The last pane is a bubble chart representing all the data in the dataset. When hovering over a bubble, information associated with the individual is presented.

All panes are tied into the filters and correspond with updated data after each selection.

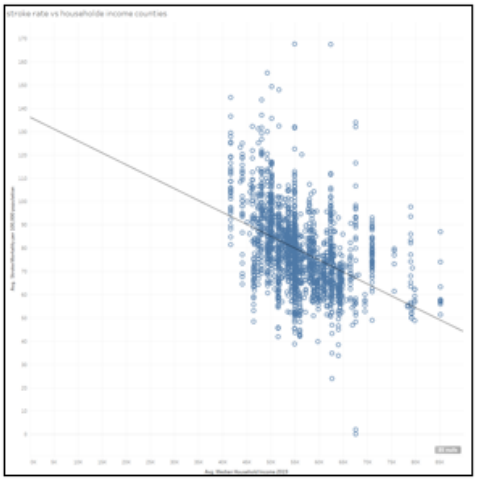


Additional data sources were used to supplement the stroke visualization effort. The data was used to create a map of stroke mortality [7] (geographic location) and associated statistics [8].

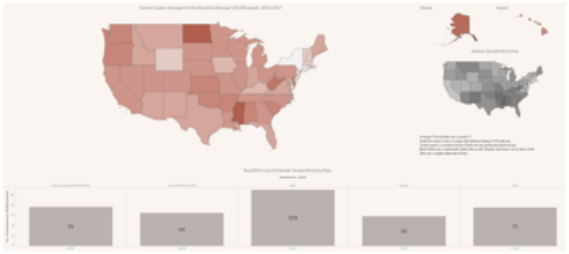
The map below combines information from both sources to create a tool to investigate stroke mortality rate along with ethnicity and average household median income.



- 3 years (2015-2017) Stroke Mortality mapped by each State
- Average stroke mortality rate is 76 per 100,000 population
- Southeast region is above average with Alabama leading at 99 deaths per 100,000



- Visible negative correlation between stroke rate and median household income
- Clear correlation with median household income for Southeast region.



- Black Stroke rate is significantly higher than overall, Hispanic and Asian is lower than overall

- Male rate is slightly higher than female.

Data Preparation for Machine Learning

Data Cleaning and Imputation

Data cleaning was conducted in Jupyter Notebook using Python.

As previously noted, the "Other" gender category was dropped from the dataset, resulting in removing 1 row of data.

In reviewing the raw data, the bmi attribute was identified as having 201 "N/A" values. This represents 3.9% of the dataset. The mean bmi value of 28.89 was used as the replacement value for the "N/A" values.

As noted above in the representation data tables, the raw dataset has a total 1,544 "Unknown" smoking status values representing 30.4% of the dataset. A closer look at the data showed 32% of the "Unknown" values were between the ages of 0-10 and 41% was between the ages of 0-15. The Centers for Disease Control and Prevention (CDC) defines a current smoker [9] as an Adult who has previously smoked 100 cigarettes in their lifetime and who currently smokes. Based on the CDC definition and the high percentage of "Unkown" values in the age range 0-10, it was originally discussed to replace those values with "never smoked". Additional research of online literature to address this issue of "Unknown" labels was conducted and it was found that "Unknown" is an accepted category. The final decision was made to leave the "Unknown" smoker status values as presented in the raw dataset.

One-Hot Encoding was used for categorical data work_type and smoking_status to be used in the linear and tree models as shown below.

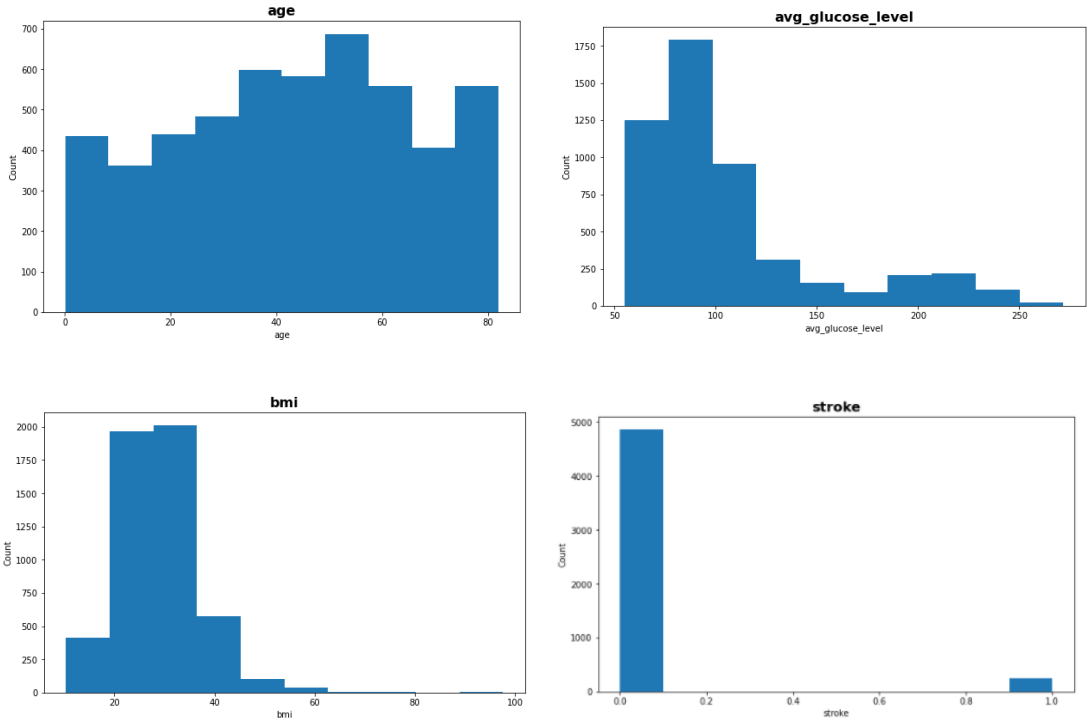
	work_type_Govt_job	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children	smoking_status_Unknown	s
0	0	0	1	0	0	0	
1	0	0	0	1	0	0	
2	0	0	1	0	0	0	
3	0	0	1	0	0	0	
4	0	0	0	1	0	0	
...	...	...	...	...	...	...	
5105	0	0	1	0	0	0	
5106	0	0	0	1	0	0	
5107	0	0	0	1	0	0	
5108	0	0	1	0	0	0	
5109	1	0	0	0	0	1	

5109 rows × 9 columns

Data Exploration

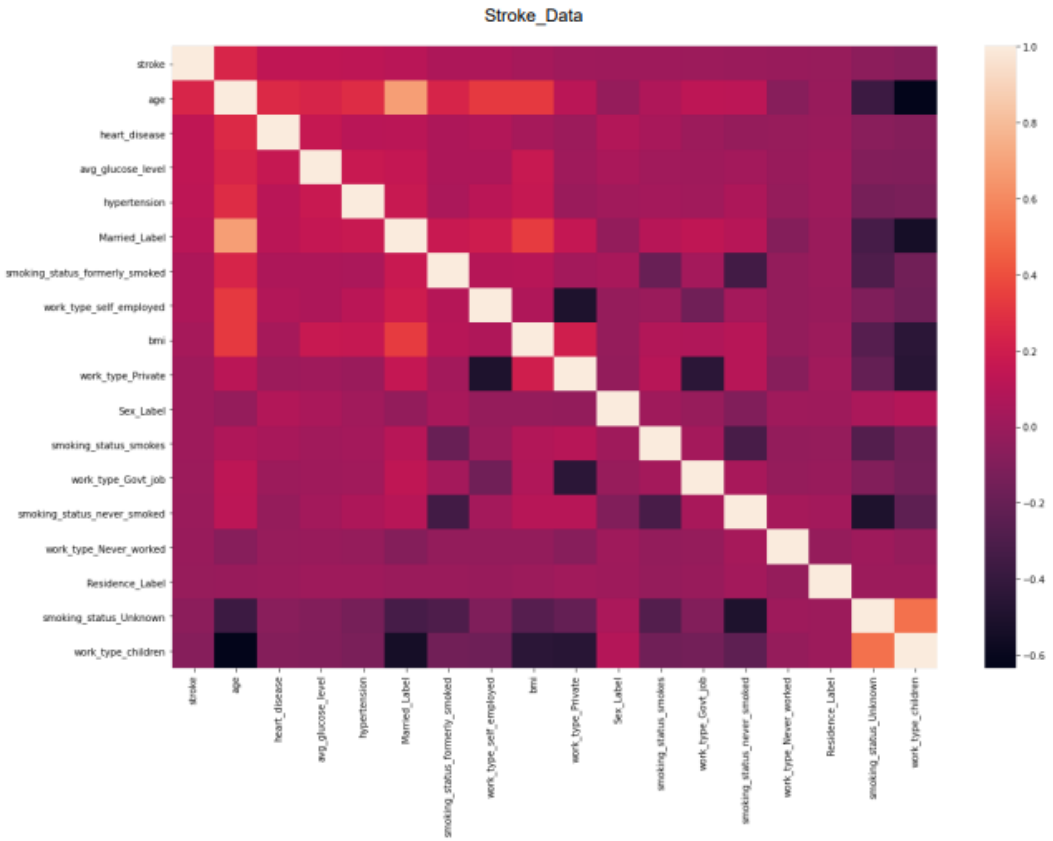
Most of the data was biased in the histograms, except for age and Residence_type. For the Yes/No questions, the data was left biased correlating to 0 which presents No as the answer to the respective question. The attributes bmi and average_glucose_level were left biased representing the lower end of their broad spectrum of data points.

Example histograms for age, bmi, average_glucose_level and stroke.



Correlation Heat Map

The correlation heat map is presented below. Values closer to zero indicate minimal to no linear relationship. The more positively correlated attributes approach 1, meaning as one attribute increases so does the other. The more negatively correlated attributes approach -1, meaning as one attribute increases the other decreases.

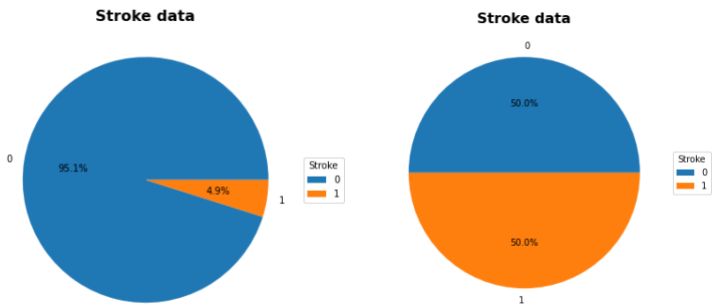


Addressing Data Bias

There is a large imbalance of stroke incidents in the dataset. To improve the model learning capabilities, bias was addressed using Synthetic Minority Oversampling Technique (SMOTE).

SMOTE utilizes k-nearest neighbor technique to create synthetic data. In this case, increase the number of stroke "Yes" values. SMOTE randomly chooses data from the stroke "Yes" values and then the respective k-nearest "No" neighbors. Synthetic "Yes" values are continually made until they closely match the "No" values. See before and after percentages below.

Stroke counts pre-SMOTE Stroke counts post-SMOTE



Machine Learning

Machine Models Evaluation

The primary objectives of the model evaluation process were to identify a model that did not overfit the data, generated the highest f1-score for 1 ("Yes" for stroke) and generated the highest recall for "Yes". The large number of 0 values ("No" for stroke), ensured a good f1-score for 0, but our objective was to identify a model that will give the best "Yes" result. That presented a challenge for the models. As noted above, SMOTE was used to help with this issue.

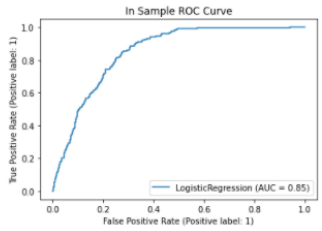
Linear Models

Models evaluated were LogisticRegression, KNeighborsClassifier and Support Vector Machines (svm). The class_weight = "balanced" parameter was set for LogisticRegression and svm. The n-neighbors = 20 parameter was set for KNeighborsClassifier.

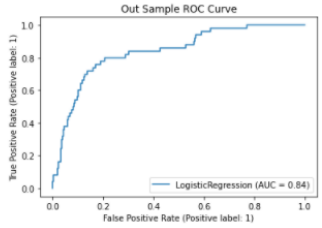
The best results for the linear model was LogisticRegression with an Out Sample f1-score of 0.23. KNeighborsClassifier and svm gave 0.16 and 0.19, respectively.

Model Evaluation Report for LogisticRegression.

```
Model Evaluation Report
In Sample Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.74      0.84      3888
           1       0.14      0.82      0.23       199

    accuracy                           0.74      4087
   macro avg       0.56      0.78      0.54      4087
weighted avg       0.95      0.74      0.81      4087

In Sample Confusion Matrix:
[[2860 1028]
 [  36  163]]
```




```
Out Sample Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.74      0.84       972
           1       0.14      0.80      0.23        50

    accuracy                           0.74      1022
   macro avg       0.56      0.77      0.54      1022
weighted avg       0.94      0.74      0.81      1022

Out Sample Confusion Matrix:
[[716 256]
 [ 10  40]]
```



Tree Models

Models evaluated were DecisionTreeClassifier, RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier and XGBClassifier. The random_state = 42 parameter was set for each tree model. The n_estimators = 1000 parameter was set for RandomForestClassifier, AdaBoostClassifier and GradientBoostingClassifier. The use_label_encoder = False parameter was set for XGBClassifier.

The best results for the tree models were AdaBoostClassifier and GradientBoostingClassifier with Out Sample f1-scores of 0.24 and 0.26, respectively. DecisionTreeClassifier, RandomForestClassifier and XGBClassifier gave 0.13, 0.16 and 0.14, respectively.

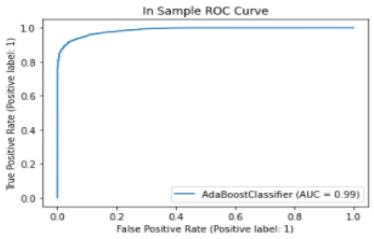
Upon further evaluation, AdaBoostClassifier and, GradientBoostingClassifier gave recall values of 0.30 and 0.20. In each case, the models had a high value of missed 1 ("Yes" for Stroke) in the Out Samples.

AdaBoostClassifier

Model Evaluation Report
In Sample Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.94	3888
1	0.94	0.93	0.94	3888
accuracy			0.94	7776
macro avg	0.94	0.94	0.94	7776
weighted avg	0.94	0.94	0.94	7776

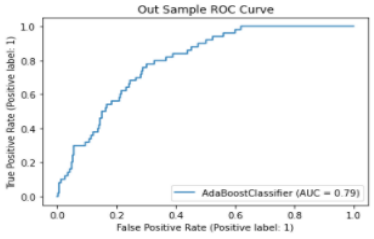
In Sample Confusion Matrix:
[[3639 249]
[253 3635]]



Out Sample Classification Report:

	precision	recall	f1-score	support
0	0.96	0.94	0.95	972
1	0.20	0.30	0.24	50
accuracy			0.91	1022
macro avg	0.58	0.62	0.60	1022
weighted avg	0.93	0.91	0.92	1022

Out Sample Confusion Matrix:
[[912 60]
[35 15]]

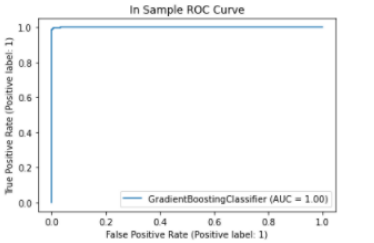


GradientBoostingClassifier

Model Evaluation Report
In Sample Classification Report:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	3888
1	1.00	0.99	0.99	3888
accuracy			0.99	7776
macro avg	0.99	0.99	0.99	7776
weighted avg	0.99	0.99	0.99	7776

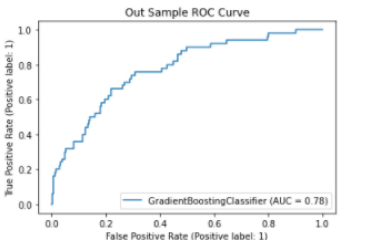
In Sample Confusion Matrix:
[[3886 2]
[51 3837]]



Out Sample Classification Report:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	972
1	0.36	0.20	0.26	50
accuracy			0.94	1022
macro avg	0.66	0.59	0.61	1022
weighted avg	0.93	0.94	0.94	1022

Out Sample Confusion Matrix:
[[954 18]
[40 10]]



Model Selection

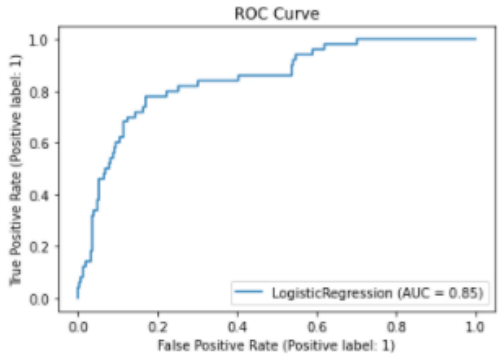
When reviewing the best of the respective Liner and Trees models, the Tree models had the best f1-scores, but extremely poor recall values. Therefore, the Liner model was selected with a little lower f1-score, but much better recall value.

Model	LogisticRegression	AdaBoostClassifier	GradientBoostingClassifier
Model Type	Linear	Tree	Tree
f1-score (1)	0.23	0.24	0.26
Recall (1)	0.80	0.30	0.20
Selection	Yes	No	No

Final Model Run

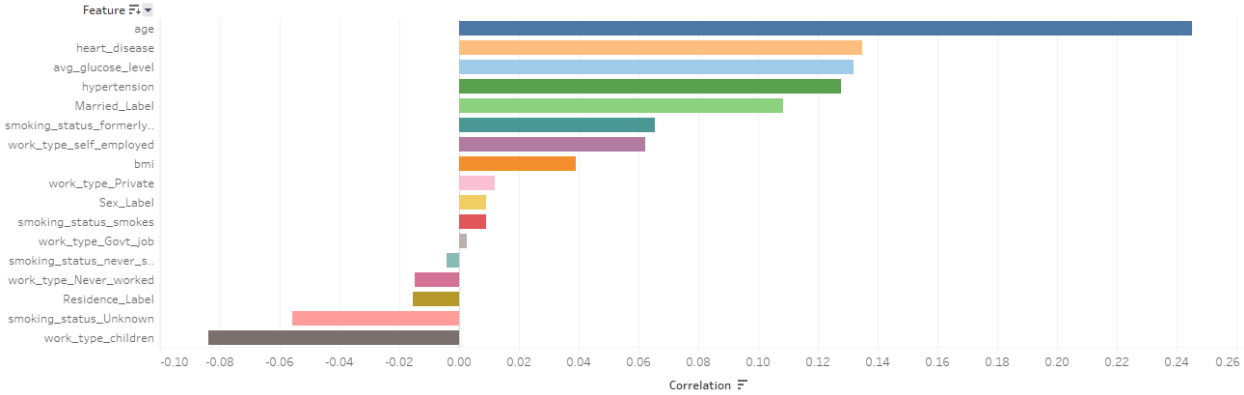
LogisticRegression was run with class_weight = balanced, max_iter = 1000 and random_state = 42.

```
Model Evaluation Report
In Sample Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.74      0.84      4860
           1       0.14      0.82      0.24       249

    accuracy                           0.74      5109
   macro avg       0.56      0.78      0.54      5109
weighted avg       0.95      0.74      0.81      5109

In Sample Confusion Matrix:
[[3583 1277]
 [  44  205]]
```



Feature Importance

Feature importance is presented below. The chart presents the assigned value of the relationship between stroke and identified attribute. Like the correlation heat map, the values closer to zero indicate minimal to no linear relationship. The more positively correlated attributes approach 1, meaning as one attribute increases so does the other. The more negatively correlated attributes approach -1, meaning as one attribute increases the other decreases.



Conclusion

Objective: To determine if a reliable model was developed, the risk factors identified by the American Stroke Association will be compared against the Features Importance table and live data will be tested.

Revisiting the Hypothesis criteria

Basis Risk Factors from American Stroke Association common to the dataset.

- High Blood Pressure
- Smoking
- Diabetes
- Obesity
- Heart Disease
- Age (cannot be controlled)
- Gender (cannot be controlled)

Hypothesis Validated

The top eleven in the Feature Importance chart with associated scores are:

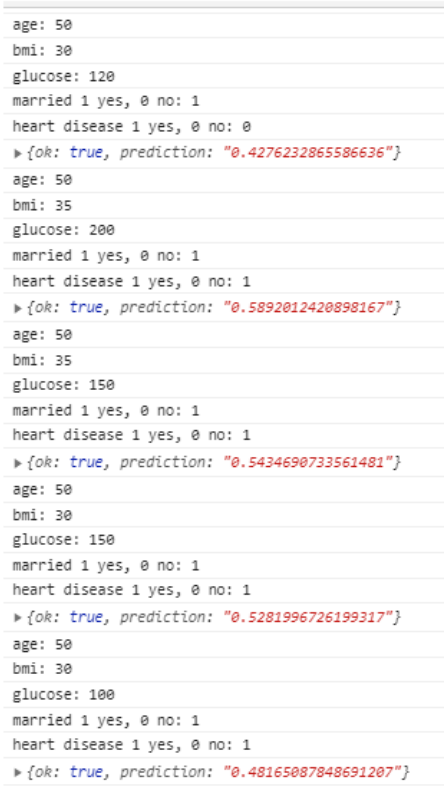
- Age – 0.2452
- Heart disease – 0.1349
- Diabetes (avg_glucose_level) – 0.1320
- High Blood Pressure (Hypertension) – 0.1279
- Married – 0.1083 - this feature was picked up by the model because of the high difference between married / not married bias in the raw data. When the data was normalized, single had the higher stroke percentages.
- Smoking (smoking_status_former – 0.0655 & smoking_status_smokes – 0.0089)
- work_type_self-employed – 0.0622 & work_type_Private – 0.0119 - an adder of stress
- Obesity (bmi) – 0.0389
- Gender – 0.0091

The risk factors from the American Stroke Association have been identified in the Feature Importance chart with positive values.

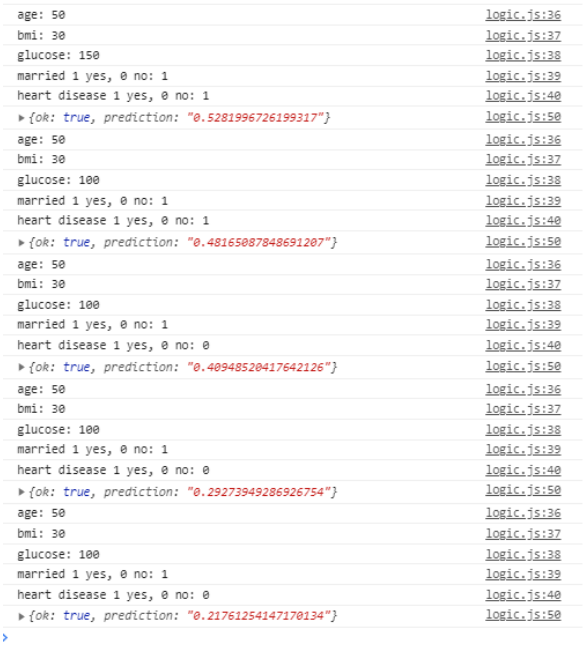
Testing of the model provides expected results as more comorbidities are added.

Trial Run Console Views

Trial 1



Trial 2



Actionable Items

This model is one of many tools which are needed to increase awareness and help reduce stroke incidents. As the noted above, the American Stroke Association states that 80% of strokes are preventable.

Actionable item

- Support stroke prevention awareness programs
  - Exercise
  - Eating correctly
  - Programs to stop smoking

Future Work

Periodic review and update of the model would be beneficial in creating a more successful tool.

References

[1] Stroke Prediction Dataset, *11 clinical features por predicting stroke events*, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[2] Ahmad FB, Cisewski JA, Miniño A, Anderson RN. Provisional Mortality Data — United States, 2020. MMWR Morb Mortal Wkly Rep 2021;70:519–522. DOI https://www.cdc.gov/mmwr/volumes/70/wr/mm7014e1.htm?s_cid=mm7014e1_w#F1_down

[3] American Stroke Association, https://www.stroke.org/en/about-stroke/types-of-stroke/ischemic-stroke-clots

[4] American Stroke Association, https://www.stroke.org/en/about-stroke/types-of-stroke/hemorrhagic-strokes-bleeds

[5] American Stroke Association, *Explaining Stroke*, pages 1-20, https://www.stroke.org/-/media/stroke-files/stroke-resource-center/brochures/explaining_stroke_brochure_6_25_19.pdf?la=en

[6] American Stroke Association, https://www.stroke.org/en/about-stroke

[7] Stroke Mortality Data Among US Adults (35+) by State…2016, Dataset in U.S. Department of Health & Human Services, https://data.world/us-hhs-gov/12ea7a13-b229-43b4-b19b-1459e9a64d3f

[8] USDA Economic Research Service, U.S. Department of Agriculture, https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

[9] Centers for Disease Control and Prevention, National Center for Health Statistics, https://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm