Stroke Prediction

An Exercise in Machine Learning and Stroke Probability Predictions

Yakov Krasnikov

Charles Philpott

Troy Youngblood

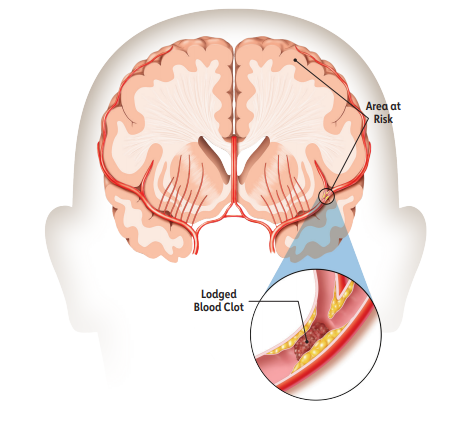
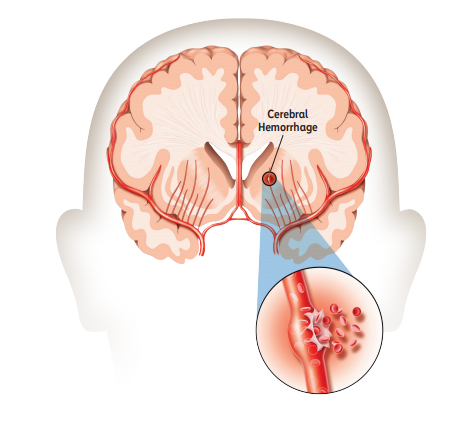
**Introduction**

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](https://www.kaggle.com/fedesoriano/stroke-prediction-dataset) [1] from Kaggle was used as the basis for the predictive model.

Cerebrovascular accidents (strokes) in 2020 were the 5th leading causeof 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. The hypothesis being tested is will the importance of the model features match the risk factors identified by the American Stroke Association.

Basis Risk Factors from American Stroke Association which can be controlled common to the dataset.

* High Blood Pressure
* Smoking
* Diabetes
* Obesity
* Heart Disease

Basis Risk Factors from American Stroke Association which cannot be controlled common to the dataset.

* Age
* Gender

**Data Source**

The data source used for the model prediction was from Kaggle [1].

The attributes with the dataset are:

* id: a unique identified 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**

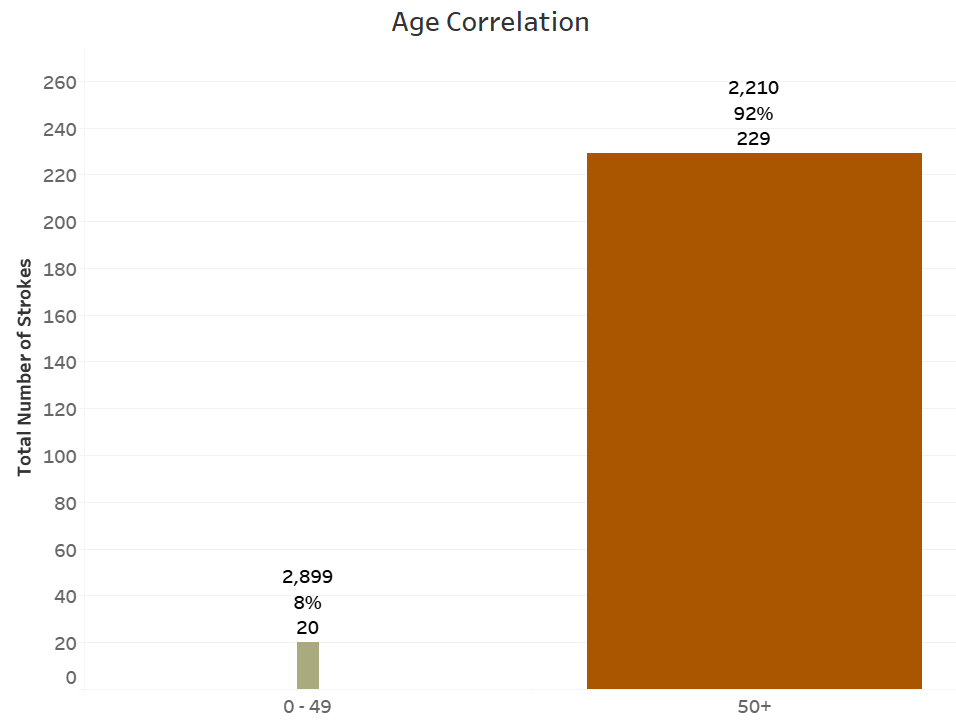
The raw dataset for machine learning consists of 5110 unique rows of information.

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 wider and darker brown bars indicate the higher normalized percentage.



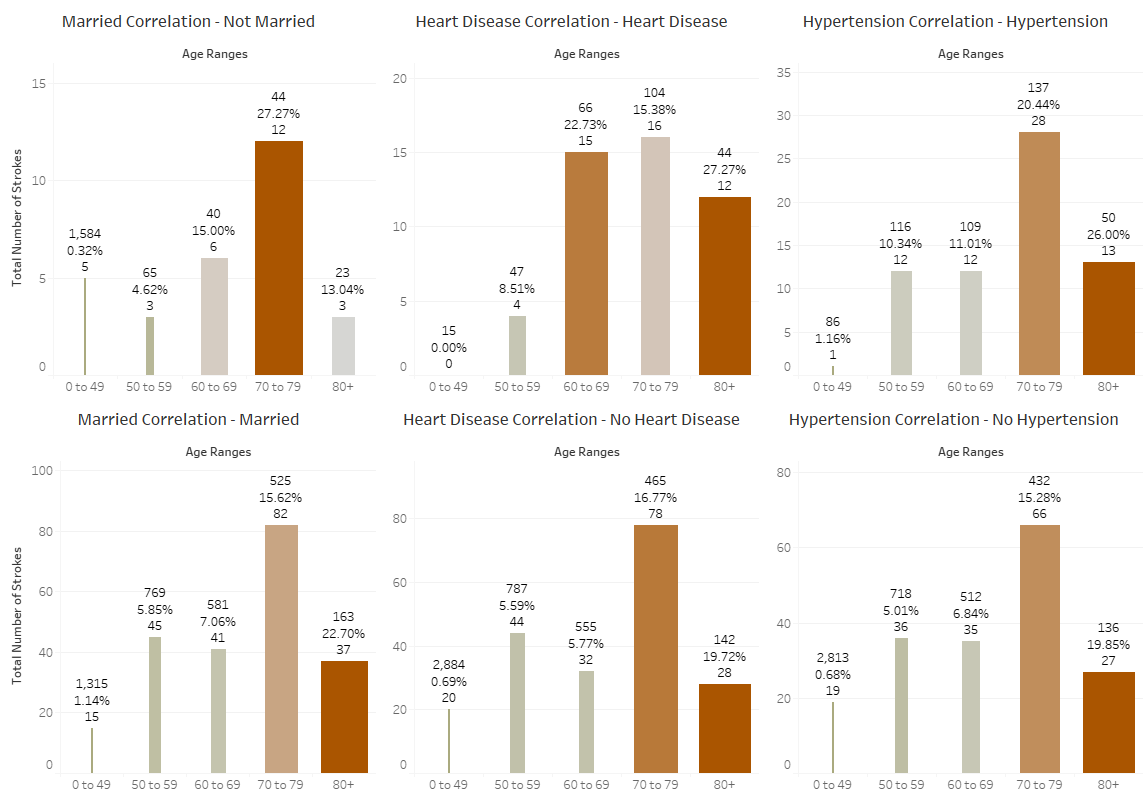
* Comorbidities Heart Disease and Hypertension, along with falling under Not Married generally have a higher percentage of strokes as age increase when the data is normalized for the specific age ranges.

The wider and darker brown bars indicate the higher normalized percentage.

The attributes are stacked for comparison purposes. The relatively equal or higher contributing factor, Not Married, Heart Disease and Hypertension are on top. The values associated with each bar in order are number of individuals in range, normalized percentage of that range and the 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.

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



* Comorbidities 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 percentage.

The BMI Categories are:

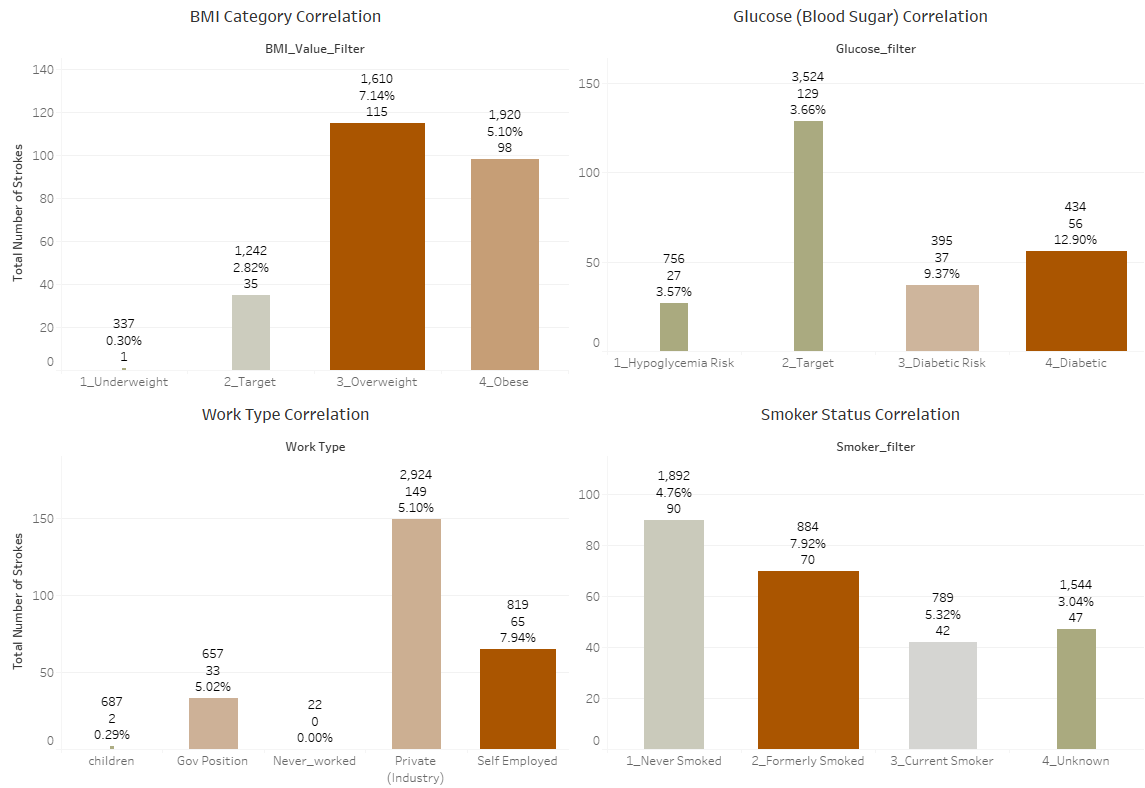
* + Underweight: BMI < 18.5
  + Target: 18.5 <= BMI < 25
  + Overweight: 25 <= BMI < 30
  + Obese: BMI > 30

The glucose data presented in the dataset is average glucose value. Depending on when a person has eaten, a glucose value will swing high and low. Therefore, to create a Glucose Category, it required a blend of Fasting, Just Eaten and Several Hours after eating ranges.

The Glucose Categories are:

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

Self Employed has the highest normalized stroke percentage for Work type.



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

Yakov - if desired - place holder to discuss supplemental data -

**Limitations of Dataset for Development of Model**

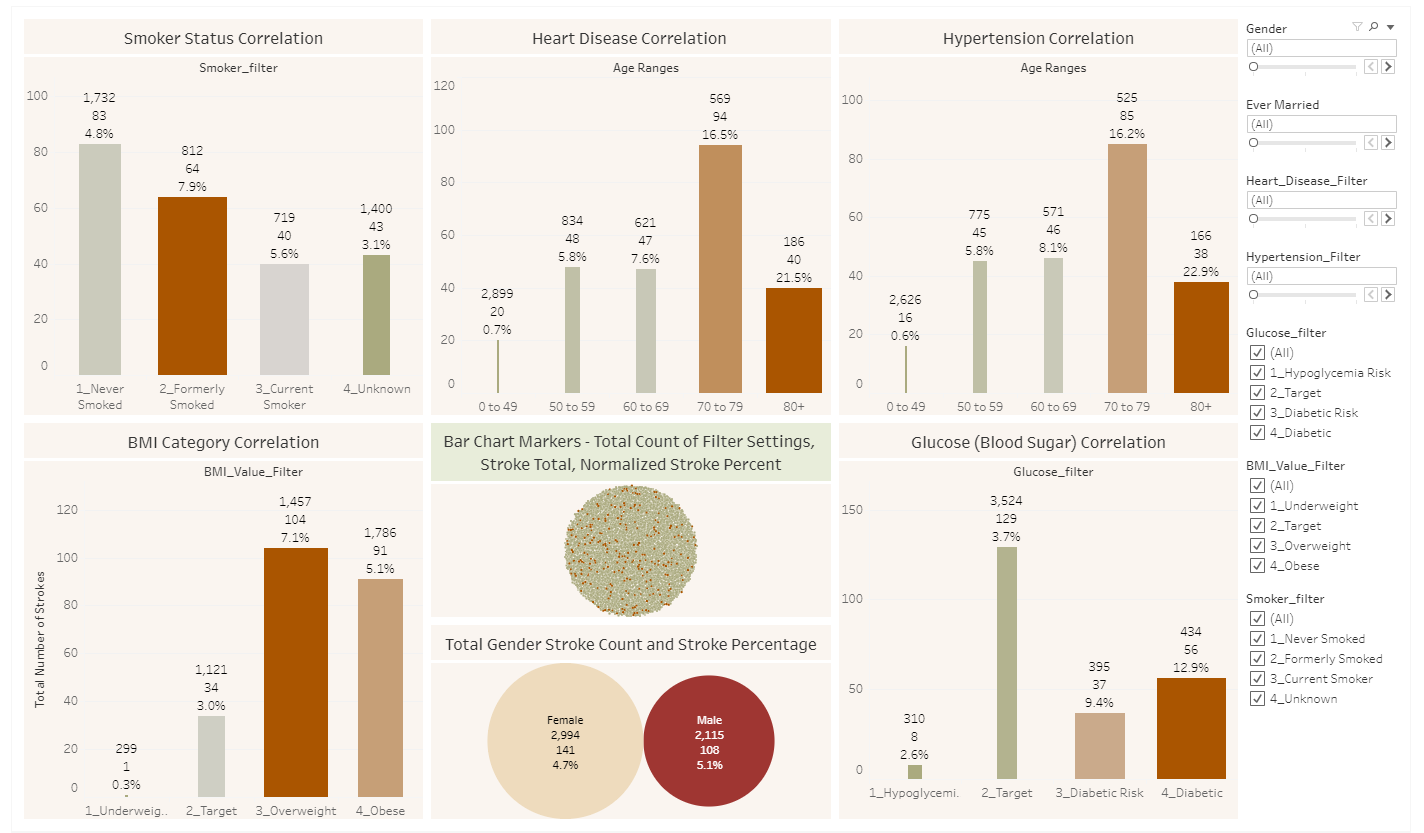
The dataset from Kaggle did not include ethnicity or economic status.

**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.



**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](https://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm) [9] as an Adult who has previously smoked 100 cigarettes in their lifetime and who currently smokes. 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 to leave the “Unknown” smoker status values as presented in the raw dataset.

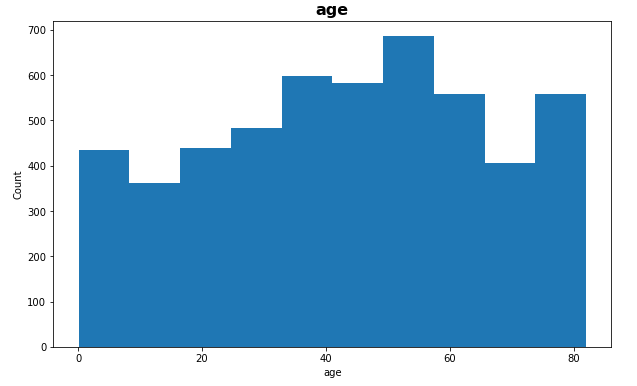
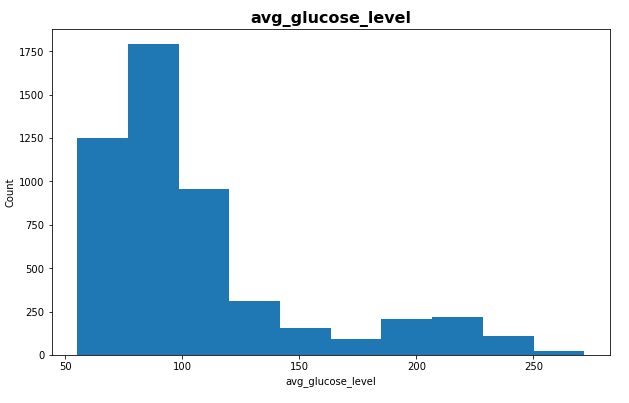
One-Hot Encoding was used for categorical data work\_type, and smoking\_status to be used in the linear and tree models.

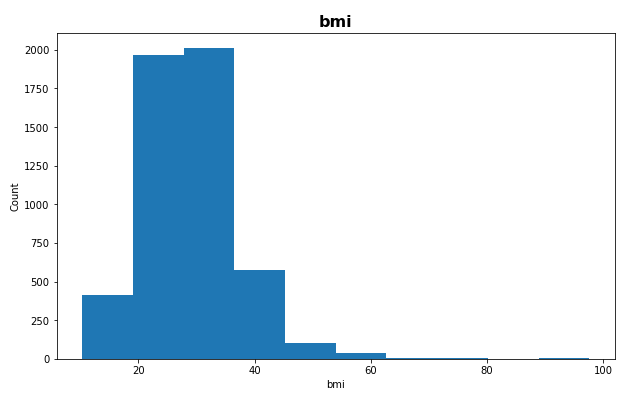
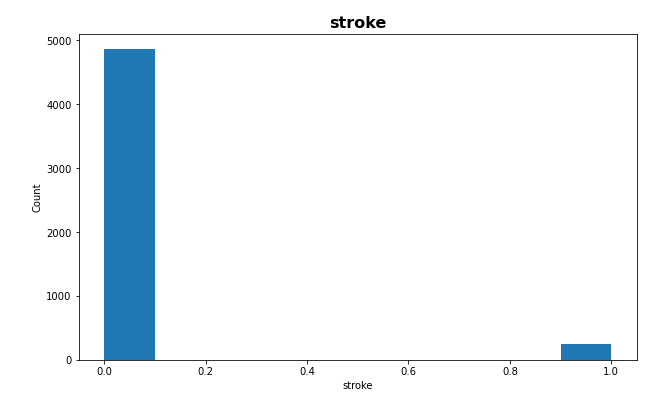


**Data Exploration**

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. BMI and average\_glucose was left biased representing the lower end of their broad spectrum of data points.

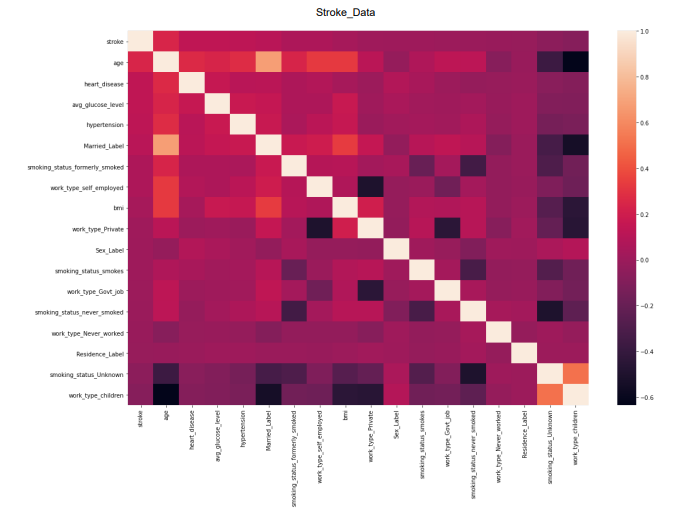
Example histograms for age, bmi and average\_glucose and stroke.

**Correlation Heat Map**

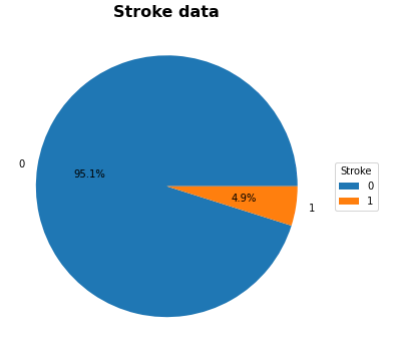
The correlation heat map is presented below.



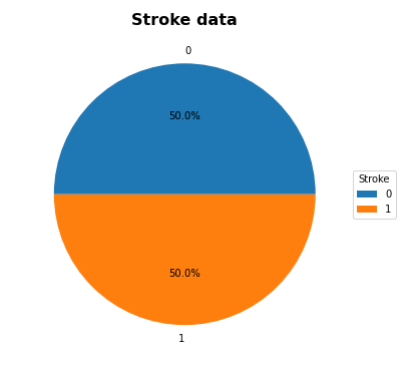
**Addressing Data Bias**

There is a large imbalance of stroke incidents in the dataset. To improve the model learning capabilities, bias was addressed using SMOTE.

Stroke Counts prior to synthesizing



Stroke data post synthesizing



**Machine Learning**

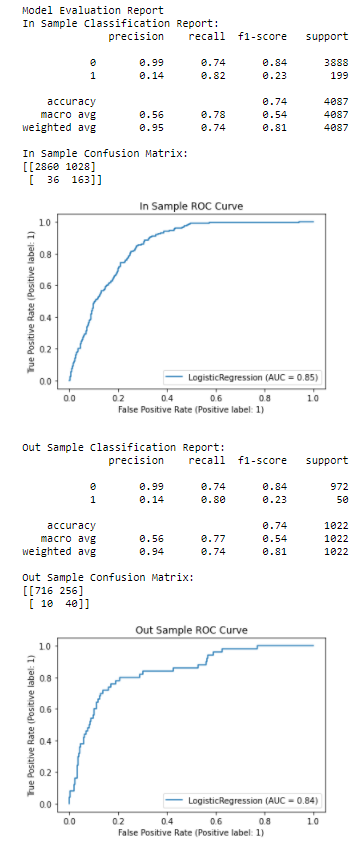
**Machine Models evaluation**

The primary objective of the model evaluation process was to identify a model that did not over fit the data, generated the highest f1-score for 1 (Yes for stroke) and generated the highest recall for 1. 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 1 results and that presented a challenge for the models.

**Linear Models**

Models evaluated were LogisticRegression, KNeighborsClassifier and Support Vector Machines (svm). The class\_weight = “balanced” was set for LogisticRegression and svm. The n-neighbors = 20 was set for KNeighborsClassifier.

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



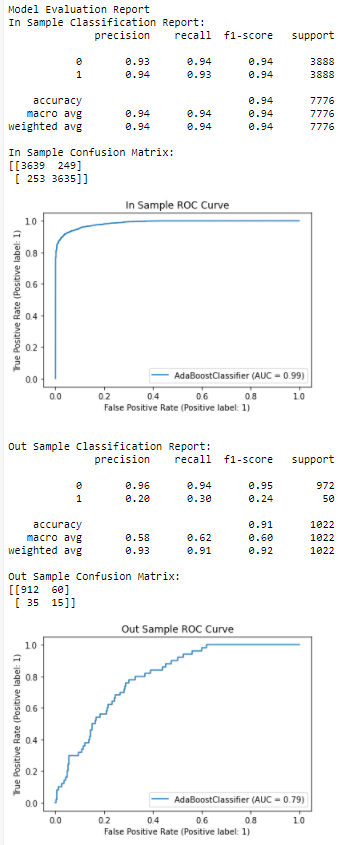
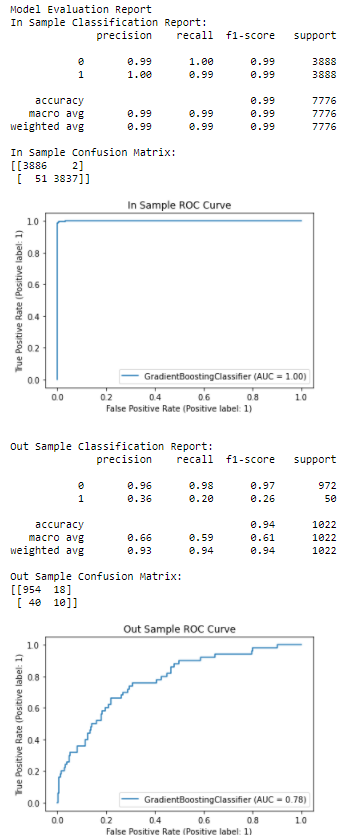
**Tree Models**

Models evaluated were DecisionTreeClassifier, RandomForestClassifier, AdaBoostClassifier, GradientBoostingClassifier and XGBClassifier. The random\_state = 42 was set for each tree model. The n\_estimators = 1000 was set for RandomForestClassifier, AdaBoostClassifier and GradientBoostingClassifier. The use\_label\_encoder = False 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 GradientBoostingClassifier

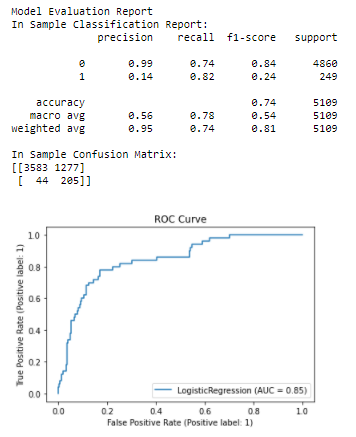
**Model Selection**

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 |

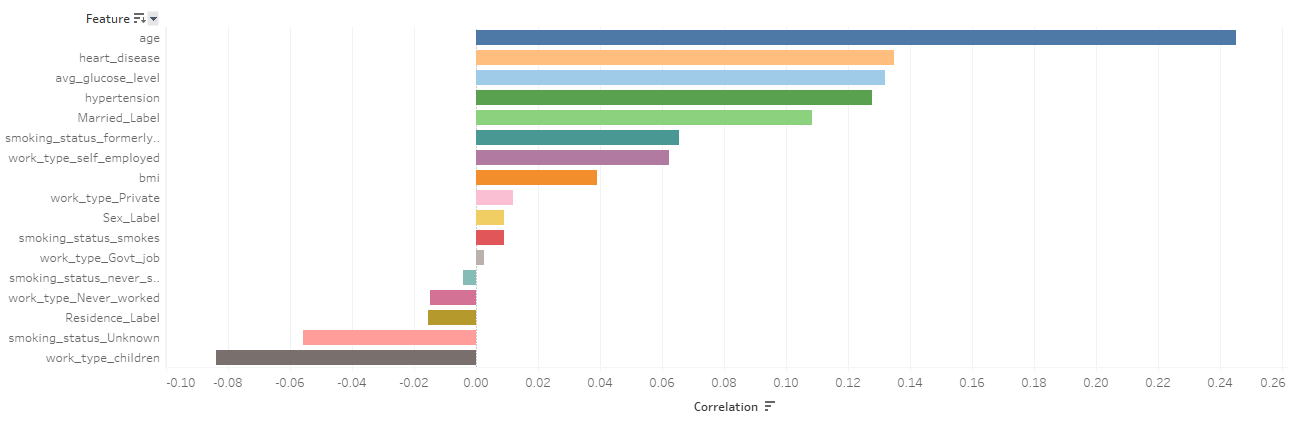
**Final Model Run**

LogisticRegression was run with class\_weight = balanced, max\_iter = 1000, and random\_state = 42.



**Feature Importance**

Feature importance is presented below.



**Conclusion**

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

Risk Factors from American Stroke Association which can be controlled common to the dataset.

* High Blood Pressure
* Smoking
* Diabetes
* Obesity
* Heart Disease

Basis Risk Factors from American Stroke Association which cannot be controlled common to the dataset.

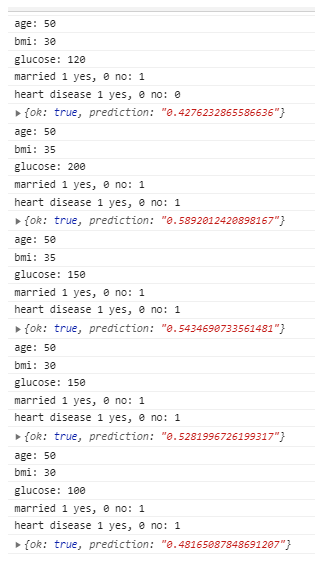
* Age
* Gender

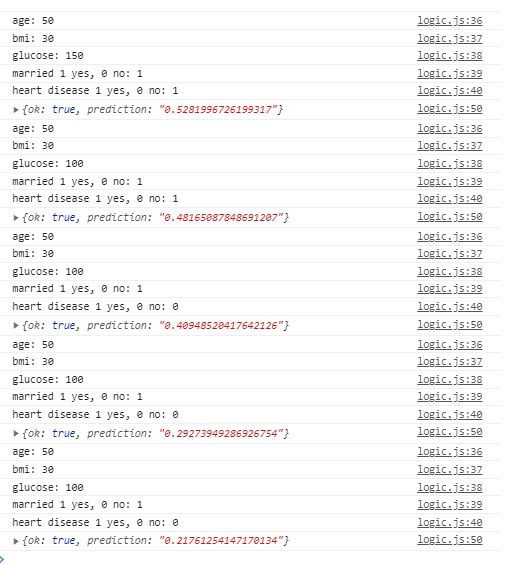
The top eleven in the Feature Importance chart with associated scores are:

* Age – 0.2452
* Heart disease – 0.1349
* avg\_glucose-level (Diabetes) – 0.1320
* Hypertension (High Blood Pressure) – 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\_status\_former – 0.0655 & smoking\_status\_smokes – 0.0089 (Smoking)
* work\_type\_self-employed – 0.0622 & work\_type\_Private – 0.0119 - an adder of stress
* bmi (Obesity) – 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.





**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 items include helping increase awareness of stroke prevention methods including exercise, eating correctly and 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>, <https://www.kaggle.com/fedesoriano>

[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: [http://dx.doi.org/10.15585/mmwr.mm7014e1external icon](http://dx.doi.org/10.15585/mmwr.mm7014e1)

[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>