Team Members (Name - NYU ID)

1. Jianglong He - jh8011
2. Renato Nunez - rn2261
3. Konner Smith - kds505
4. Atharva Bhagwat - acb9244

Problem

The World Health Organization (WHO) has found that stroke is the 2nd leading cause of death worldwide. Each year, 15 million people suffer a stroke and approximately a third of the occurrences are fatal. In addition to the emotional toll that having a stroke usually entails, there is also a substantial economic cost. In the United States alone, it is estimated that Americans miss out on $15 billion in lost wages due to a stroke annually.

There are multiple known factors that put a person at higher risk of suffering a stroke. Some (family history, age, gender, race) cannot be changed. Fortunately, others, such as exercise habits and other lifestyle factors, are within a patient's control.

Given the seriousness of stroke, it is crucial that predictive models be conservative. To be successful, a model must predict all true positives and have no false negatives. Analyses must be evaluated based on how many positive observations were correctly predicted.

Therefore, two of the most important attributes of a successful data science analysis are:

- Provides insight into treatable risk factors
- Correctly predicts positive observations while minimizing false negatives

Related Work

Some of the most popular and accurate models used in stroke prediction problems include decision trees, logistic regression, Naive Bayes, SVM, and KNN classifiers. Out of these models, one study concludes that Naive Bayes is the most accurate classifier at 82% accuracy.

There are also a lot of publications in the Kaggle platform where people use various models mentioned above along with data oversampling strategy. However, most publications perform sampling on the dataset before doing a train-test split. This will lead to data leakage and also change of the true distribution of the test dataset.

Performing oversampling on the dataset before splitting will create duplicates of positive samples and hence when one performs the split the test dataset is likely to have duplicate samples that are in the training dataset. There are oversampling techniques such as SMOTE where it does not duplicate the minority samples instead it generates new samples that are close to the original. But even wil SMOTE, one will be using the information from the whole dataset to create the new samples and hence there will still be data leakage issues. Most publications achieve more than 90% F1-scores by falsely using the sampling techniques. Hence, these scores are not reflecting the true performance of the trained model when it comes to new unseen data points.

Sources:
https://www.datasciencecentral.com/profiles/blogs/stroke-prediction-using-data-analytics-and-m achine-learning

https://thesai.org/Downloads/Volume12No6/Paper_62-Analyzing_the_Performance_of_Stroke_ Prediction.pdf

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/code

Data Science Objectives & Key Results

Understanding the Data:

- Understand the relationship between features with respect to the target variable.
- Understand the actual medical parameters which lead to a stroke and check if the same relation is followed by the dataset.

Modeling:

- Train supervised binary classification models using various algorithms to optimize the recall score of stroke class.
- Reduce the number of features and model complexity by removing correlating features.
- Treat this task as a “Anomaly Detection” problem and work with single class models to work around the imbalance in the dataset.

Success Criteria:

- Optimizing recall score for stroke class. As this problem is in the healthcare domain, it is crucial to minimize the number of false negative detections. So a higher recall is very important in determining the performance of a model. Miss classifying a positive observation can even lead to death. Further tests can be

implemented to check the accuracy of the positive prediction but missing it entirely should be penalised.

Data Exploration

We use the Stroke Prediction Dataset that is available on Kaggle. The dataset is confidential and collected in a real world setting. Dataset is provided in tabular format where each row represents a data point and each column represents a specific feature. There are a total 5110 data points with 11 features and 1 target label column. The dataset contains both categorical and continuous features and the target label is a categorical variable with 2 unique values. So the task can be narrowed down to a supervised binary classification problem.

Since this dataset is a real world dataset, it is expected that it will not be perfect to use without some data preprocessing. Some data preprocessing steps include converting categorical data into numerical values, handling missing values, and checking for outliers. Techniques to handle the missing values include filling them with the mean of the feature, dropping the data points with the missing value, or finding data points with other similar features and filling the missing value with that.

We also spot that the standard deviation varies from feature to feature, so some data normalization is needed. One of the major problems with this dataset is class imbalance. There are 249 data points with class “1” and 4861 with class “0”. Approaches to solve this problem are data oversampling, data undersampling, and transforming the task into an anomaly detection task. Data oversampling duplicates the data points with the minority class label. On the other hand, data undersampling drops data points with a majority class label. Using these two techniques we try to balance out the class labels. However, one downside for using data oversampling and undersampling is that it might change the true distribution of the data. Thus we need to be very careful to only perform oversampling and undersampling only on the training dataset not the testing dataset.

We also treat this task as an anomaly detection task where the model learns to detect “anomalies” or “outliers”. This technique is also called one-class classification. We fit the model only using the data points from the majority class and treat the minority class as “anomalies” for the model to avoid. During the training phase, the model learns to capture the distribution of the majority class. During the inference phase, any new incoming data points that fall outside the learned distribution will be classified as an outliner.

We use precision, recall, and f1 score as the evaluation metric because they are sensitive to false positives and false negatives. Using a plain accuracy as metric is not ideal here because of the class imbalance problem. The model can learn to always predict the majority class and that will give it a good accuracy. However, it is obvious that the model has not learned anything. Using precision, recall, and f1 score the model metric can be penalized harder if they make any false prediction on the data points with the minority class. We want to put more attention to the recall score of the model, especially the recall score of positive samples (having stroke) because it will be very bad if our model predicts a false negative as it might cause delay in treatment of someone who is likely to have a heart stroke.

Key Activities

Data Acquisition

Dataset is sourced from Kaggle:
https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Real World Context of Features

Gender

While men have more strokes, women generally have more severe strokes. Currently strokes kill twice as many women as breast cancer. A variety of factors play a part in the increased death rate for women such as pregnancy and greater life expectancy (due to the fact that stroke risk also increases with age).

Age

Stroke risk increases with age as arteries become narrower and harder, every decade over 55 doubles the probability of stroke.

Hypertension

Hypertension, high blood pressure, can be the result of unhealthy lifestyle choices and/or family history. It is when your blood is consistently applying too much force or pressure against your artery walls, damaging your blood vessels as well as your arteries. Hypertension is commonly considered the most significant risk factor for a stroke.

Heart Disease

Heart disease and stroke share many of the same risk factors such as hypertension. Heart disease or a heart attack occur when your arteries are blocked.

Marriage

The impact of marriage on stroke risk is highly dependent on other contextual factors. For example a healthy marriage can decrease stroke risk and a past marriage that ended in divorce

can increase stroke risk. This primarily relates to the significance of mental health in determining stroke risk.

Employment
Similar to marriage, employment can have an impact on mental health which can in turn increase or decrease stroke risk. Long work hours or highly stressful occupations will increase stroke risk, however the categories covered in the work_type feature of our dataset do not provide that information.

Geography (residence_type)
In the US there is an 11 state region known as the “stroke belt” where stroke risk is highest. The stroke belt includes Mississippi, Tennessee, Louisiana, Kentucky, Georgia, North Carolina, Alabama, South Carolina, Arkansas, Indiana and Virginia. All of these states are located in the southern region of the US and have a large amount of rural communities.

Average Glucose Level
A high average glucose level can be an indicator of diabetes or pre diabetes. A person with diabetes is 1.5 times more likely to have a stroke.

Body Mass Index (BMI)
Being overweight can increase stroke risk at any age and also leads to high blood pressure which is one of the main causes of stroke.

Smoking
Not only does smoking increase the likelihood of stroke but also increases the severity making it twice as likely for a smoker to die of stroke. The risk also increases rapidly the more someone smokes.

Exploratory Data Analysis

The dataset has 5110 observations and 10 features and 1 target variable. The 10 **features are:**

1) gender
2) age
3) hypertension
4) heart_disease
5) ever_married
6) work_type
7) Residence_type
8) avg_glucose_level
9) bmi
10) smoking_status

The target variable is: stroke

Out of these 10 features, medically most critical features are: hypertension, heart_disease, smoking_status, bmi, and age.

Analysis of Target Variable:

```
Count:
0    4861
1     249
Name: stroke, dtype: int64

Percentage:
0    95.13
1     4.87
Name: stroke, dtype: float64
```

95% of total observations are negative stroke samples. This set of observations are highly skewed.

Analysis of Categorical Features:

Categorical features are:

1) **gender:** Has 3 unique values. Male, Female, and other. There is only one observation with gender as 'other'. Therefore we delete this observation treating it as an outlier.

```
Percentage distribution for gender:
stroke  gender
0       Female    58.69
        Male      41.29
        Other      0.02
1       Female    56.63
        Male      43.37
```

   We see equal distribution of observations with respect to gender and stroke.

2) **hypertension:** Has 2 unique values. 0 and 1. 0 stands for no hypertension while 1 stands for hypertension.

```
Percentage distribution for hypertension:
stroke  hypertension
0       0             91.11
        1              8.89
1       0             73.49
        1             26.51
```

Hypertension is one of the important features according to medical sources. But, we see no such trend in our dataset. About 73.5% of positive stroke samples have no hypertension.

3) **heart_disease:** Has 2 unique values. 0 and 1. 0 stands for no heart disease while 1 stands for heart disease.

```
Percentage distribution for heart_disease:
stroke  heart_disease
0       0              95.29
        1               4.71
1       0              81.12
        1              18.88
```

Heart disease is one of the important features according to medical sources. But, we see no such trend in our dataset. About 81% of positive stroke samples have no history of heart disease.

4) **ever_married:** Has 2 unique values. Yes and No.

```
Percentage distribution for ever_married:
stroke  ever_married
0       Yes            64.45
        No             35.55
1       Yes            88.35
        No             11.65
```

We do see that about 88% of stroke samples are married. But, 65% of no stroke samples are also married meaning a positive value for feature ever_married does not mean much.

5) **work_type:** Has 5 unique values. Private, Self-employed, Govt_job, children, and Never_worked.

```
Percentage distribution for work_type:
stroke  work_type
0       Private          57.11
        Self-employed    15.51
        children         14.09
        Govt_job         12.84
        Never_worked      0.45
1       Private          59.84
        Self-employed    26.10
        Govt_job         13.25
        children          0.80
```

We see the feature work_type has the same distribution for positive and negative stroke samples.

6) **Residence_type:** Has 2 unique values. Rural and Urban.

```
Percentage distribution for Residence_type:
stroke  Residence_type
0       Urban             50.63
        Rural             49.37
1       Urban             54.22
        Rural             45.78
```

Residence_type has near equal distribution for both values of stroke variable.

7) **smoking_status:** Has 4 unique values. formerly smoked, never smoked, smokes, and Unknown.

```
Percentage distribution for smoking_status:
stroke  smoking_status
0       never smoked       37.07
        Unknown            30.80
        formerly smoked    16.77
        smokes             15.37
1       never smoked       36.14
        formerly smoked    28.11
        Unknown            18.88
        smokes             16.87
```

Smoking status is one of the important features according to medical sources. But similar to heart disease and hypertension, this dataset does not show that trend. We see that for both values of stroke variable, the distribution of smoking status is the same.

Just by looking at the categorical features, we cannot find any certain relationship between target variable and the feature values.

To further investigate the relationships between categorical variables, we conducted chi-square testing between all 21 possible pairs of categorical variables. Of these pairs, 17 were determined to be statistically significant (Note: Statistical significance was determined by the standard test of p-value <= 5%). Several of the relationships support consensus knowledge in the medical community (E.g. The relationships between hypertension and heart disease, gender and heart disease).

The full list of statistically significant relationships are their p-values is below:

Feature #1	Feature #2	P-Value
ever_married	work_type	0.00000000000000000000%
work_type	smoking_status	0.00000000000000000000%
ever_married	smoking_status	0.00000000000000000000%

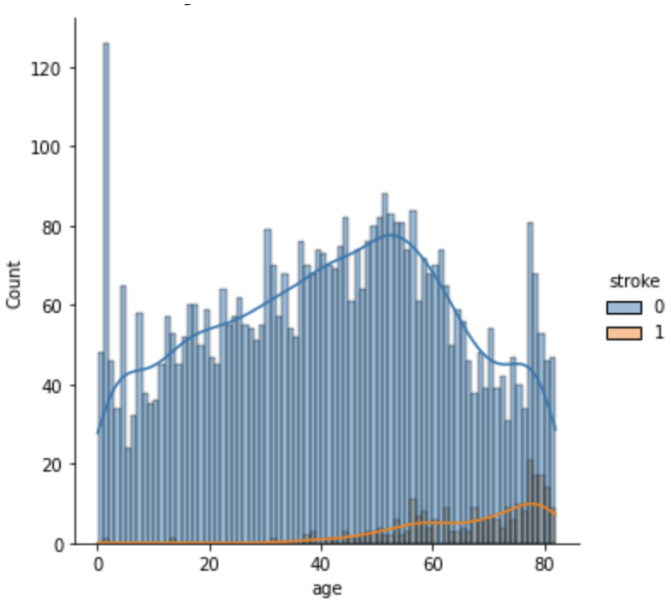
hypertension	ever_married	0.000000000000000000000%
hypertension	work_type	0.000000000000000000000%
hypertension	smoking_status	0.000000000000000000022%
heart_disease	ever_married	0.000000000000044277708%
heart_disease	work_type	0.000000000001623361832%
hypertension	heart_disease	0.000000000002238628302%
gender	smoking_status	0.000000022273124272157%
heart_disease	smoking_status	0.000001051058070589 61%
gender	heart_disease	0.000001347141153357 43%
gender	work_type	0.000014316073809600 30%

Analysis of Continuous Features:

	age	avg_glucose_level	bmi
count	5110.000000	5110.000000	4909.000000
mean	43.226614	106.147677	28.893237
std	22.612647	45.283560	7.854067
min	0.080000	55.120000	10.300000
25%	25.000000	77.245000	23.500000
50%	45.000000	91.885000	28.100000
75%	61.000000	114.090000	33.100000
max	82.000000	271.740000	97.600000

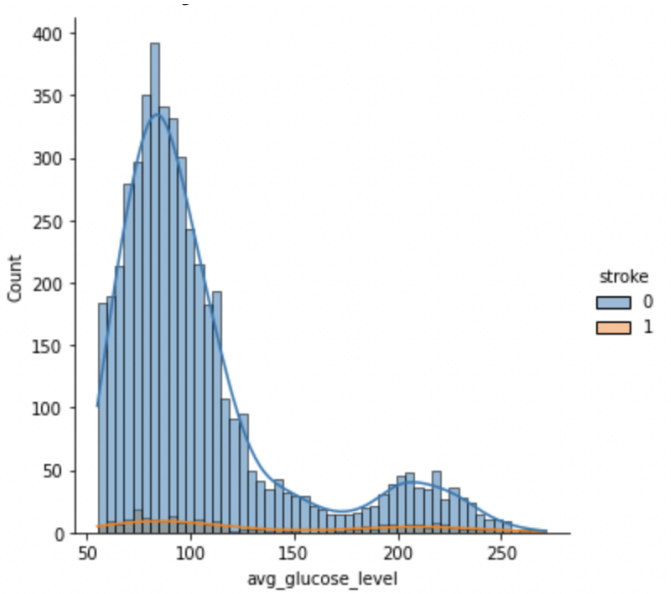
Continuous features are:

1) **age:**



We observe that the distribution of positive stroke samples is higher for ages greater than 40. A positive stroke sample has a 97% chance of having age greater than 40. As the number of negative stroke samples are way higher, this trend is only 4% of all the observations. Age is also one of the important features as per medical sources, while we cannot base our predictions on this feature alone, age has high importance.

2) **avg_glucose_level:**

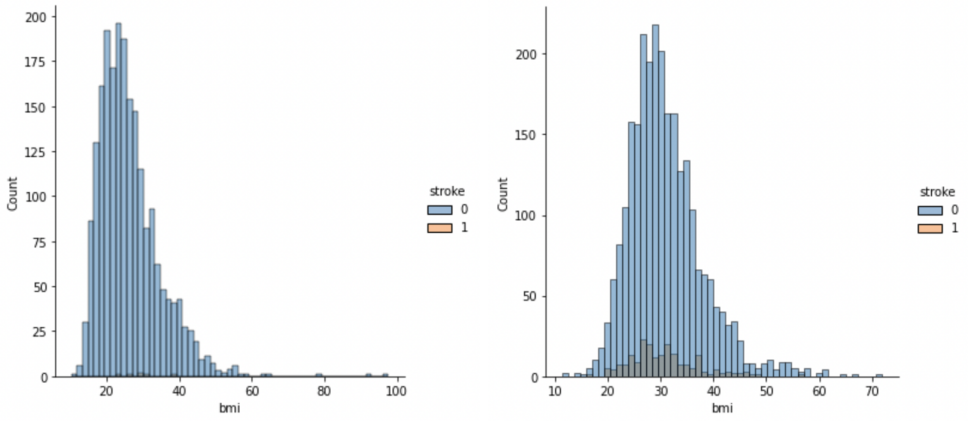


According to medical sources, average glucose level less than 90 causes hypoglycemia and higher than 150 leads to hyperglycemia causing damage to

the pancreatitis. These ‘abnormal’ values can cause strokes in patients with other conditions.

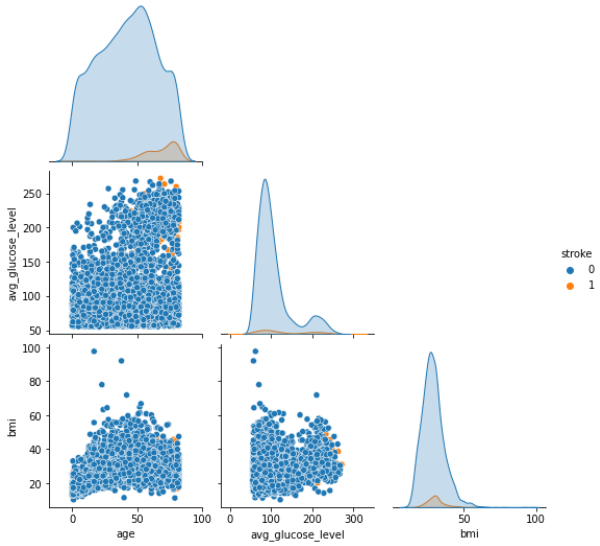
We observe that the distribution of positive stroke samples is higher for avg_glucose_level less than 90 or avg_glucose_level greater than 150. A positive stroke sample has a 71.5% chance of having ‘abnormal’ average glucose level. As the number of negative stroke samples are way higher, this trend is only 5.7% of all the observations.

3) **bmi:**



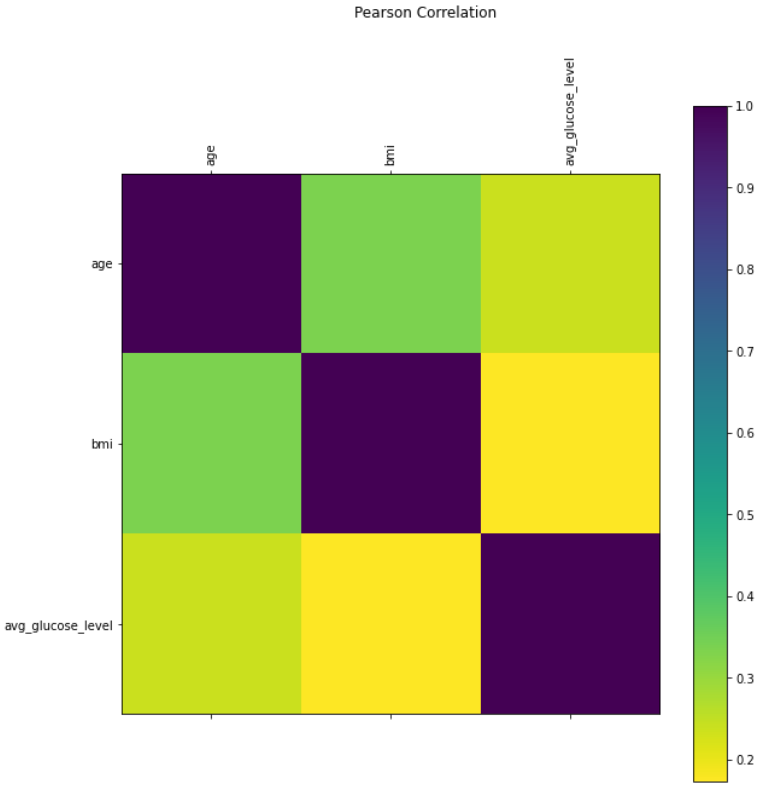
BMI is one of the important features according to medical sources. A BMI in range 18.5 - 24.5 is considered normal. Above plots indicate bmi values for age less than 40(left) and greater than 40(right). We see that the distribution of BMI is the same for both sets.

Pairplot (Continuous Features):



In the above plot we do not see distinct clusters for positive and negative stroke samples. The positive samples really are ‘anomalies’.

Correlation (Continuous Features):



There are no strong correlations between the continuous variables. We see only a weak correlation between ‘age’ and ‘bmi’.

Preprocessing:

About 3.9% of observations have bmi value missing. We filled the missing values by grouping the observations based on age and calculating the mean.

We also removed a row with the gender value equal to ‘other’ as there was just one such observation.

To prepare our dataset for our predictive models, we converted several categorical features to integers:

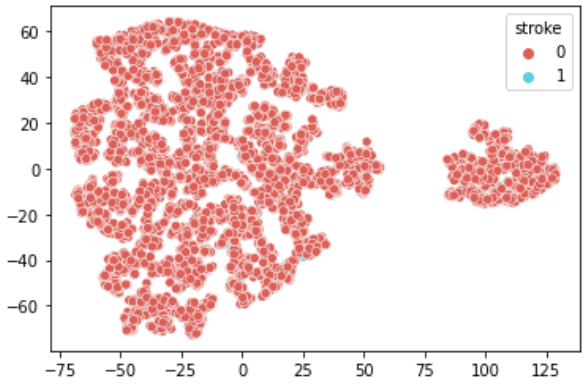
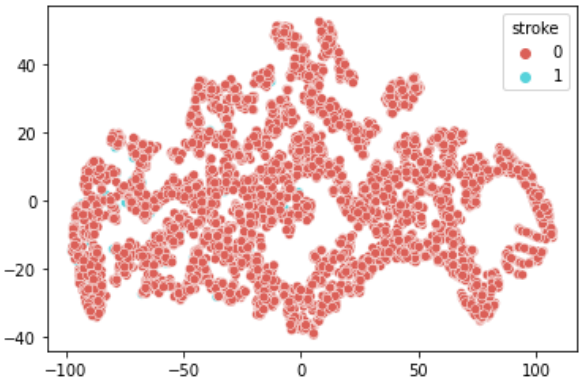
- Gender: Binarized (0 = Female, 1 = Male)
- Ever_married: Binarized (0 = No, 1 = Yes)
- Hypertension: Binarized (0 = No, 1 = Yes)
- Heart_disease: Binarized (0 = No, 1 = Yes)
- Work_type: Converted the five work types to 0, 1, 2, 3, and 4
- Residence_type: Binarized (0 = Rural, 1 = Urban)
- Smoking_status: Converted the four statuses to 0, 1, 2, and 3

Creating subset with medically sound features:

We will be working with two datasets. One will be the entire dataset and the other will be a subset with only the medically sound features. ie: hypertension, heart_disease, smoking_status, bmi, and age.

T-SNE Plot:

Let's look at the T-SNE plot for both the sets.

Complete Set	Medical Set
	

Even with reducing the dimensions of the dataset based on similarity of the points, we see no distinct clusters.

Data preparation:

Stratification

We use stratified train test split with test size equal to 10% of total number of observations.

Feature Scaling

Standardize features by removing the mean and scaling to unit variance.

Tabulating results: (Optimizing Recall)

Model	Dataset	Class	Accuracy (Test)	Recall	Precision	F1 Score
		No Stroke	-	0.418	0.971	0.584

Single Class SVM	Complete	Stroke	-	0.76	0.063	0.116
		Overall	0.434	0.589	0.517	0.35
	Medical	No Stroke	-	0.344	0.982	0.509
		Stroke	-	0.88	0.065	0.12
		Overall	0.37	0.612	0.524	0.314
Isolation Forest	Complete	No Stroke	-	0.414	0.962	0.578
		Stroke	-	0.68	0.056	0.104
		Overall	0.44	0.547	0.509	0.341
	Medical	No Stroke	-	0.514	0.977	0.674
		Stroke	-	0.76	0.075	0.139
		Overall	0.515	0.637	0.526	0.405
SVM	Complete (Undersampled)	No Stroke	-	0.335	0.994	0.502
		Stroke	-	0.96	0.069	0.129
		Overall	0.366	0.648	0.532	0.316
	Medical (Undersampled)	No Stroke	-	0.599	0.98	0.743
		Stroke	-	0.76	0.089	0.159
		Overall	0.629	0.68	0.534	0.451
	Complete (Oversampled)	No Stroke	-	0.451	0.986	0.619
		Stroke	-	0.88	0.076	0.14
		Overall	0.472	0.66	0.531	0.38
	Medical (Oversampled)	No Stroke	-	0.7	0.983	0.817
		Stroke	-	0.76	0.115	0.2
		Overall	0.703	0.73	0.549	0.508
	Complete	No Stroke	-	0.726	0.983	0.836
		Stroke	-	0.76	0.125	0.215

Naive Bayes		Overall	0.728	0.743	0.554	0.525
	Medical	No Stroke	-	0.774	0.978	0.864
		Stroke	-	0.66	0.13	0.218
		Overall	0.768	0.717	0.554	0.541
	Uncorrelated	No Stroke	-	0.623	0.987	0.768
		Stroke	-	0.84	0.104	0.185
		Overall	0.639	0.734	0.546	0.185
K-Nearest Neighbors	Complete	No Stroke	-	0.93	0.97	0.95
		Stroke	-	0.36	0.20	0.26
		Overall	0.68	0.64	0.58	0.60
	Medical	No Stroke	-	0.76	0.99	0.86
		Stroke	-	0.88	0.16	0.27
		Overall	0.70	0.82	0.58	0.57
Logistic Regression	Complete	No Stroke	-	0.84	0.98	0.91
		Stroke	-	0.68	0.18	0.29
		Overall	0.83	0.76	0.58	0.60
	Medical	No Stroke	-	0.87	0.98	0.92
		Stroke	-	0.64	0.21	0.31
		Overall	0.86	0.76	0.59	0.62
Multi Layer Perceptron	Complete	No Stroke	-	0.88	0.98	0.93
		Stroke	-	0.64	0.21	0.32
		Overall	0.86	0.76	0.59	0.63
	Medical	No Stroke	-	0.88	0.98	0.93
		Stroke	-	0.64	0.22	0.33
		Overall	0.87	0.76	0.60	0.63

Conclusion:

Key factors for stroke detection in real world

Real world risk factors can be broadly divided into two categories; treatable and untreatable. Treatable factors include high blood pressure (hypertension), heart disease, diabetes, smoking habits, a sedentary lifestyle, and obesity. Untreatable factors include gender, age, genetics, and race.

Key factors spotted by EDA and model

While no distinct clusters are visible for pairs of features, models have slightly better performance on the subset with medically sourced features. With more observations, models will be able to better classify test samples.

With the trained Logistic Regression model, we get insight on which features did the model focus on by looking at the model coefficients. The coefficients of the model are following:

```
[('work_type', -0.0588163926305361 4),
 ('ever_married', -0.05333671617313357),
 ('bmi', -0.0006027325216923931),
 ('gender', 0.012112108176674841),
 ('Residence_type', 0.056124273100585105),
 ('smoking_status', 0.06499904847009667),
 ('heart_disease', 0.07137364242049679),
 ('avg_glucose_level', 0.14153720322172755),
 ('hypertension', 0.1438739473926299),
 ('age', 1.820741495786744)]
```

As we can see, the Logistic Regression model puts the majority of weights on the “age” feature in order to predict a sample to be “stroke”. Followed by a similar amount of weights on [“hypertension”, “avg_glucose_level”], further followed by [“heart_disease”, “smoking_status”, “residence_type”]. Features that are more important in predicting “no stroke” are [“work_type”, “ever_married”, “bmi”]. The model did focus on features that are claimed to be more important in detecting stroke when it comes to real world scenarios. Features like “blood pressure (hypertension)”, “heart disease”, “smoking habits” and “age” are heavily weighted by the trained model in deciding whether a data sample is likely to have heart stroke.

How can we possibly improve in the future

The biggest challenge we face in this project is the imbalance of the dataset. If we can gather more positive (stroke) data points to balance out the dataset then our models should perform much better.

References:

https://www.stroke.org.uk/sites/default/files/smoking_and_the_risk_of_stroke.pdf
https://www.verywellhealth.com/being-overweight-and-stroke-risk-3146345
https://www.diabetes.org/diabetes/complications/stroke
http://www.thepreventioncenter.com/cardiovascular-disease/stroke-belt/
https://www.verywellhealth.com/your-marriage-affects-your-chances-of-having-a-stroke-3145876
https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/stroke-risk-factors-and-prevention#risk-factors-of-stroke
https://www.cdc.gov/stroke/women.htm
https://www.stroke.org.uk/what-is-stroke/are-you-at-risk-of-stroke
https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/risk-factors-for-stroke
https://pubmed.ncbi.nlm.nih.gov/10863872/
https://pubmed.ncbi.nlm.nih.gov/29127059/
https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/risk-factors-for-stroke