Project Report: Identifying a patients' susceptibility to stroke

By Aman Oberoi

Index

Executive Summary

Stroke is a relatively preventable medical emergency. Data of previous patients who have suffered from stroke can help us **identify patterns of health conditions that can lead to stroke**. In this project, I analyze, clean and enhance the data provided and run several different machine learning models to develop an accurate classifier of patients based on their likeliness of getting a stroke.

Analyzing the dataset by examining the distributions of the different features, and by calculating the correlations between the predictor features and the relevant stroke values shows that the **age, average glucose levels and the heart disease history** of a person are the **most important factors** that can help determine their susceptibility to a stroke. The dataset had few missing values for the BMI of patients but I handled that by replacing them with the median values for BMI since the data points could prove to be useful.

After the data was processed through a pipeline detailed later, running OLS regression on the data to caclulcate p-values of variables with respect to the stroke labels showed that **the gender, type of work and the smoking habits of a patient had high p-values.** This means that **these features are not the strongest indicators** of whether a person has had a stroke or not. To combat these high p-values, the machine learning models were tested on simplified data (through dimensionality reduction).

Since we are trying to predict whether a person has had or could get a stroke based on their medical information, we need high accuracy as well as a **good recall score** so that whenever a person is highly susceptible to strokes we diagnose them correctly and give them the appropriate healthcare. The **best overall model** that has good accuracy and good recall is the logistic regression model with a modified threshold value that has a **test accuracy of 0.8003 (80.03%) and a recall of 0.722 (72%)**.

Introduction

The dataset provided includes detailed information on the health profile of patients with factors such as their gender, age, hypertension, glucose level, etc. However, these **attributes are not always sufficient** to predict if someone is at risk of getting a stroke. Other factors that can lead to a stroke are a sedentary lifestyle, dehydration and poor diet. Individuals such as young adults who are otherwise healthy and have normal blood glucose levels, don't smoke and have no heart disease or hypertension can also get a stroke if they have prolonged states of dehydration, strong medication or poor diet. **To account for these factors, we would need information on additional features such as the amount of exercise a person gets every week and their diet and drinking habits.**

Thus it is important to stress that while the attributes provided can be a good indicator of a person's susceptibility to strokes, they do not take into account small fluctuations in a person's health such as a few weeks of poor diet. Birth defects in blood vessels are also known to be a leading cause of strokes in people - such defects can go unnoticed for years and people might not even know they have the defects till they get a stroke. Another challenge is that there are lots of different types of strokes which can occur due to different combinations of factors and this type of data on patients suffering from different strokes can confuse our machine learning models.

Methodology

1) **Data inspection:** In my project, I first transformed all data into numeric types (even categorical variables) in order to plot the feature distributions as well as to plot the correlation matrix of the data.
2) **Data pipeline:** I then re-imported the data to run through my data processing pipeline. I first identified the variables with null values (only 'bmi') and then converted the categorical ever_married and Residence_type variables to numeric 0s or 1s (through label encoding) because each of those features only have two possible values. Finally, for the numeric variables, I imputed the variable (bmi) with missing values with the median value. I also added a new cross feature that computes the product of the BMI and glucose_level for each data point (patient). Finally all the numerical variables were mean standardized using sklearn's StandardScaler.
3) **P-value analysis and dimensionality reduction:** I then carried out OLS regression on the processed data to compute the p-values of all the different features and identify any redundant, noisy variables. Followed by the results of this OLS regression, I reduced the dimensionality of the processed data by running principal component analysis on it using sklearn's decomposition.
4) **Preparing training and testing data:** After dimensionality reduction, I split the processed, dimensionality reduced data into train and test splits. I also artificially balanced the data using SMOTE to have a 2:1 ratio of non stroke and stroke patients so that my models have better recall and are better at identifying cases of patients that are susceptible to strokes.
5) **Logistic regression:** I conducted a small test by running three logistic regression models on processed data that had not been dimensionally reduced and on processed data that had been dimensionally reduced. The accuracy of all three models was better when trained on the dimensionally reduced data. I also identified the best of the three models using 10-fold cross validation and plotted its accuracy, recall, precision f1 score and confusion matrix. I was not

satisfied with the recall of my model so I reduced the threshold of the same model to 0.4 and got a better recall but lower accuracy.

6) **Ensemble model:** I also trained four different random forest classifiers on the processed data with different hyperparameters and plotted the stats of the best one (after running 10-fold cross validation).
7) **Neural network:** Finally, I trained neural networks on my processed data. I created 3 different neural network architectures by changing parameters such as the activation functions of the hidden layers and the number of neurons in each layer. I conducted 10-fold cross validation to find the neural network with the best accuracy and then printed out its statistics.

Results

Data statistics

- I plotted the distribution of all the different variables to get a better understanding of the data (figure 1).

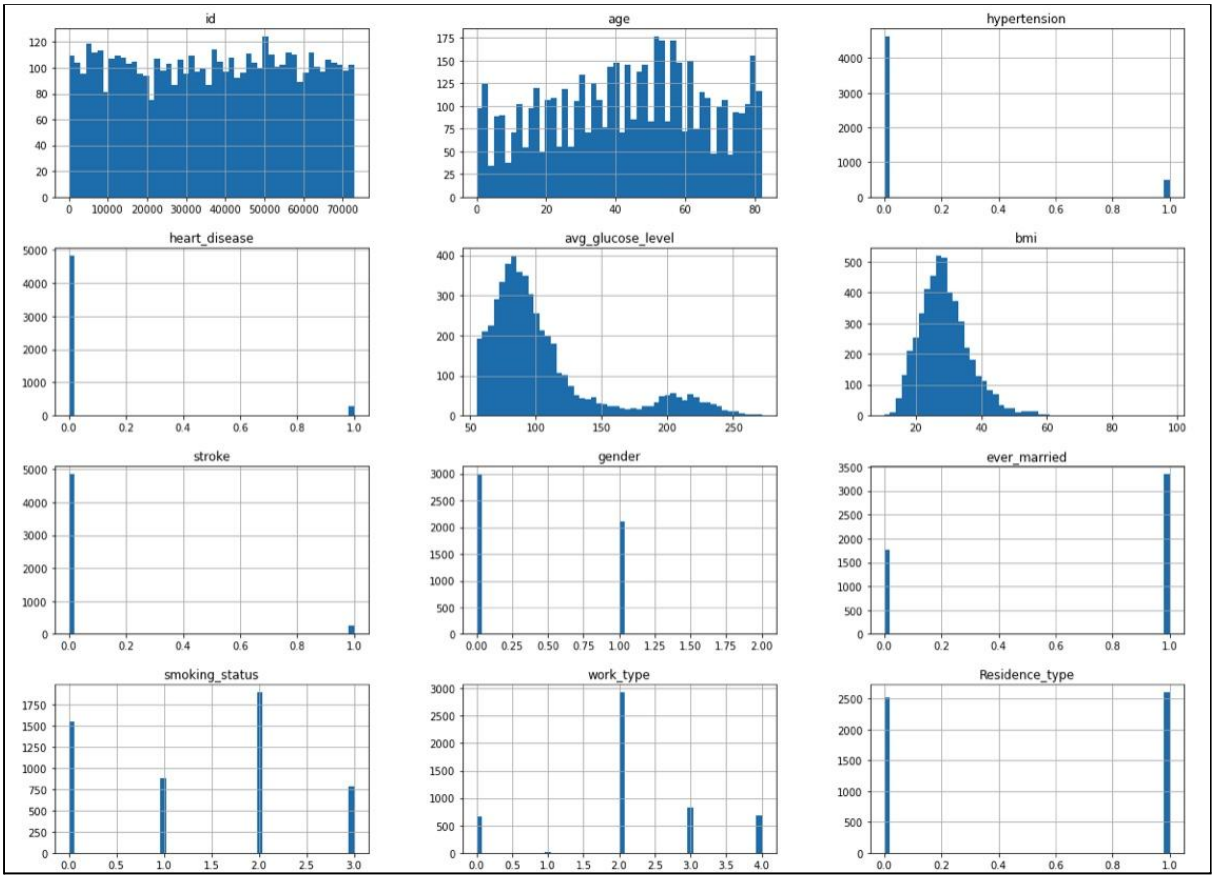


Figure 1: distributions of all the features of the dataset

- I analyzed the stroke data provided to identify any important correlations. Plotting a correlation matrix of data shows that some of the features most highly correlated with a person's stroke history are the person's age, their average glucose level and their heart disease history (figure 2).

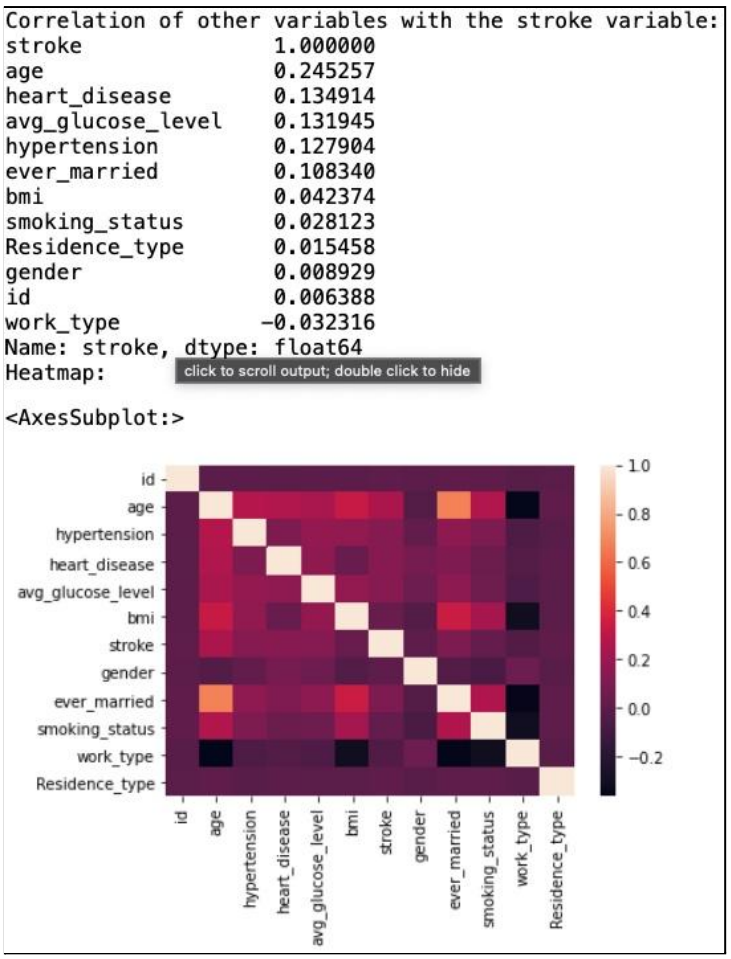


Figure 2: Correlation values with respect to stroke variable and correlation matrix of all the data

- On initial correlation analysis, it is observed that age is the attribute most highly correlated to the stroke attribute, followed by heart disease, avg glucose level and hypertension. This indicates that people who have had heart disease, are older, have higher glucose levels or have hypertension are at a higher risk of having a stroke than others. While it looks like smoking, gender, residence type are not correlated with strokes, they are actually categorical variables which were converted into numeric types to calculate the correlation matrix. These variables might turn out to be important.
- A flaw in the dataset was that around 5% of the data points were patients that have had a stroke (figure 3). This is an unfair representation of the population. It is predicted by some websites that around 1 in 4 people over the age of 25 have a stroke at some point in their lives. To make our dataset more balanced

and to train better models, we would need more data points of people who have had a stroke, or artificially balance the data.

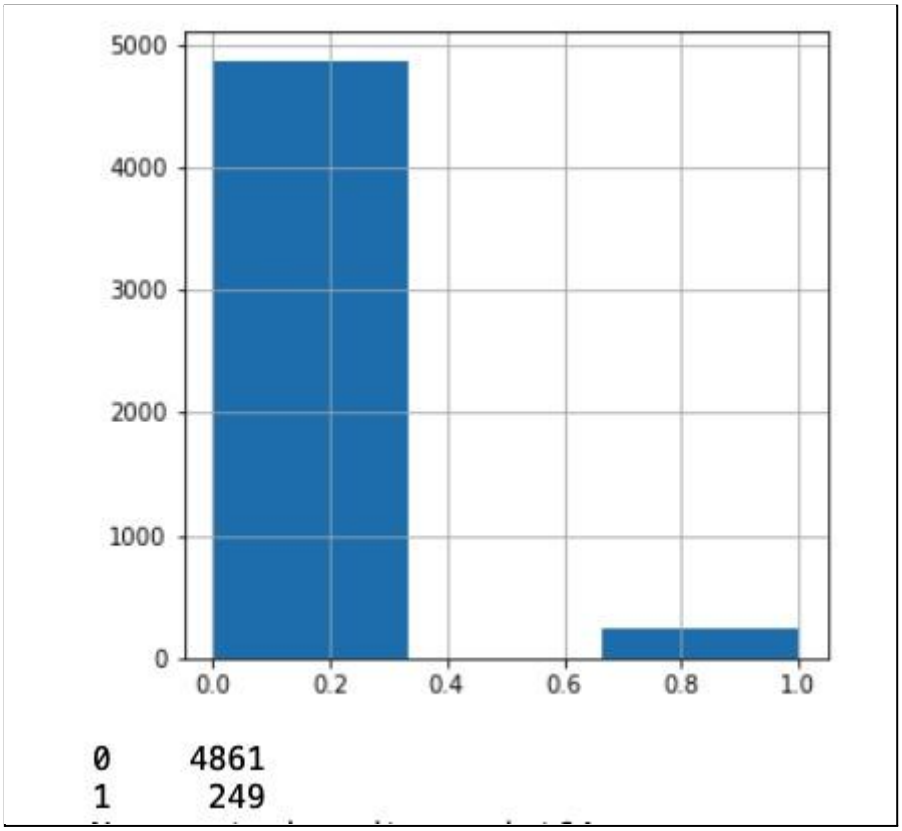
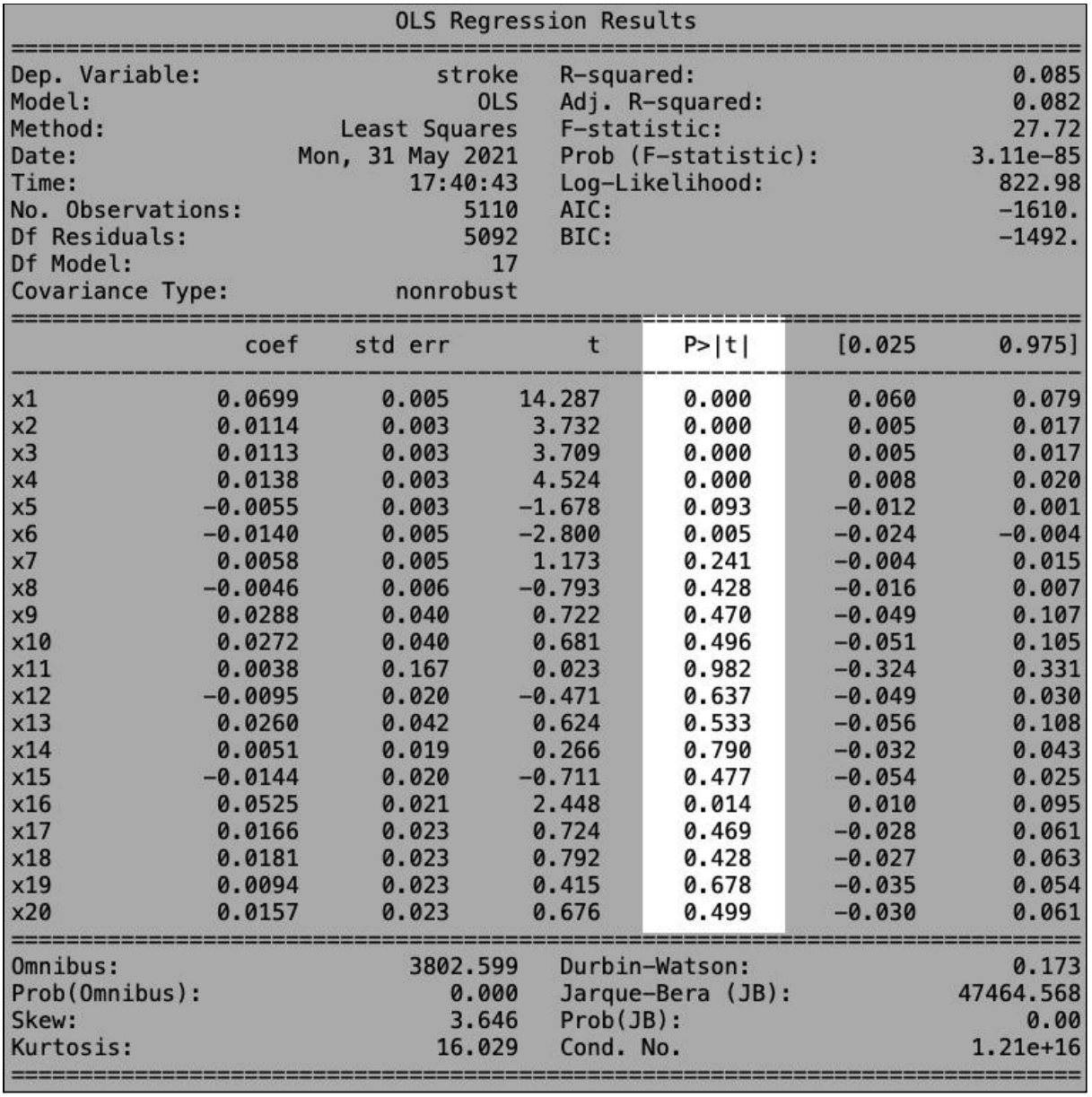


Figure 3: distribution of the stroke variable - a very small percentage of patients have had a stroke

Data processing

- After data inspection, I processed the data by sending it through a pipeline that mean normalized numerical variables, one-hot encoded the relevant categorical variables, imputed the missing values of the BMI feature and added an augmented feature (BMI * glucose_level).
- Running OLS regression on the processed data showed that a lot of features had high p-values which indicates that the features are not statistically significant and will just add noise to the data when models are being trained on them. (figure 4)



OLS Regression Results

Dep. Variable:	stroke	R-squared:	0.085
Model:	OLS	Adj. R-squared:	0.082
Method:	Least Squares	F-statistic:	27.72
Date:	Mon, 31 May 2021	Prob (F-statistic):	3.11e-85
Time:	17:40:43	Log-Likelihood:	822.98
No. Observations:	5110	AIC:	-1610.
Df Residuals:	5092	BIC:	-1492.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.0699	0.005	14.287	0.000	0.060	0.079
x2	0.0114	0.003	3.732	0.000	0.005	0.017
x3	0.0113	0.003	3.709	0.000	0.005	0.017
x4	0.0138	0.003	4.524	0.000	0.008	0.020
x5	-0.0055	0.003	-1.678	0.093	-0.012	0.001
x6	-0.0140	0.005	-2.800	0.005	-0.024	-0.004
x7	0.0058	0.005	1.173	0.241	-0.004	0.015
x8	-0.0046	0.006	-0.793	0.428	-0.016	0.007
x9	0.0288	0.040	0.722	0.470	-0.049	0.107
x10	0.0272	0.040	0.681	0.496	-0.051	0.105
x11	0.0038	0.167	0.023	0.982	-0.324	0.331
x12	-0.0095	0.020	-0.471	0.637	-0.049	0.030
x13	0.0260	0.042	0.624	0.533	-0.056	0.108
x14	0.0051	0.019	0.266	0.790	-0.032	0.043
x15	-0.0144	0.020	-0.711	0.477	-0.054	0.025
x16	0.0525	0.021	2.448	0.014	0.010	0.095
x17	0.0166	0.023	0.724	0.469	-0.028	0.061
x18	0.0181	0.023	0.792	0.428	-0.027	0.063
x19	0.0094	0.023	0.415	0.678	-0.035	0.054
x20	0.0157	0.023	0.676	0.499	-0.030	0.061

Omnibus:	3802.599	Durbin-Watson:	0.173
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47464.568
Skew:	3.646	Prob(JB):	0.00
Kurtosis:	16.029	Cond. No.	1.21e+16

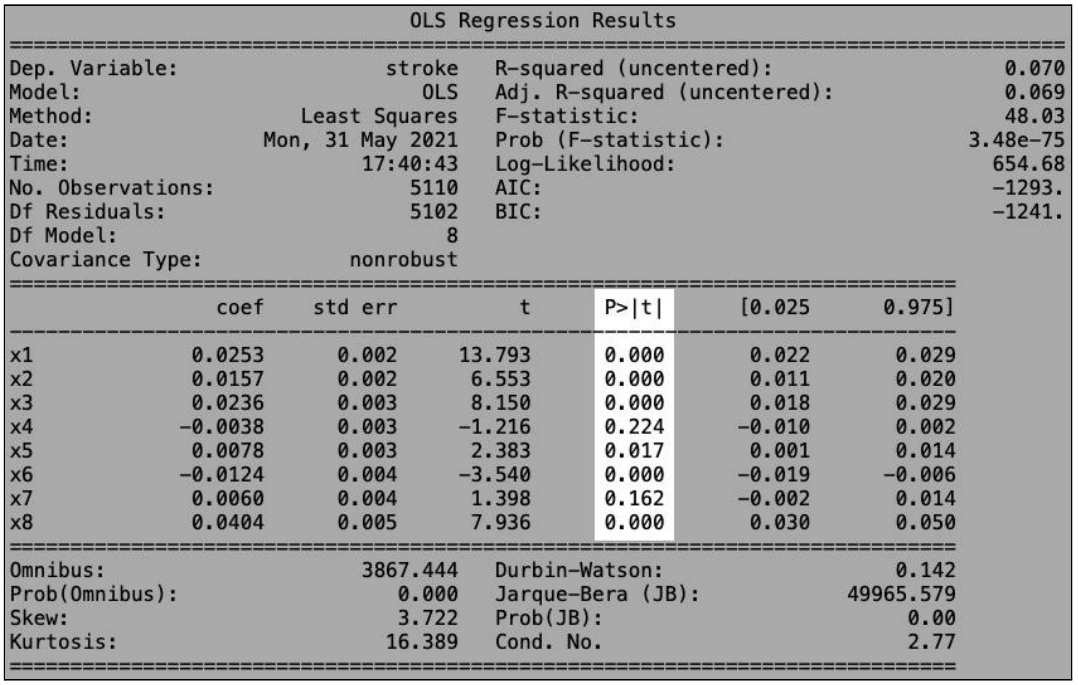
Figure 4: results of OLS regression on processed data - the highlight column indicates the p values

(Note that in the table in figure 4 the following are the variables: 1 - age 2 - hypertension 3 - heart rate 4 - average glucose level 5 - BMI 6 - ever married 7 - residence type 8 - BMI x glucose (cross term) 9 - female 10 - male 11 - other 12 - govt job 13 - never worked 14 - private work 15 - self employed 16 - child 17 - unknown smoking status 18 - formerly smoking 19 - never smoked 20 - smokes)

- From the data in figure 4 it looks like the 'other' gender type, private work and the 'never smoked' smoking type are variables with high p values which means that the stroke labels (whether someone has had a stroke or not) does not depend much on these variables. This makes intuitive sense for never smoked -

as although a person doesn't smoke, they are still susceptible to strokes due to other factors.

- Running OLS regression with the dimensions of the data reduced to 8 using PCA shows that the p-values reduce significantly (figure 5). Unfortunately, the variables can no longer be interpreted meaningfully.



OLS Regression Results

Dep. Variable:	stroke	R-squared (uncentered):	0.070
Model:	OLS	Adj. R-squared (uncentered):	0.069
Method:	Least Squares	F-statistic:	48.03
Date:	Mon, 31 May 2021	Prob (F-statistic):	3.48e-75
Time:	17:40:43	Log-Likelihood:	654.68
No. Observations:	5110	AIC:	-1293.
Df Residuals:	5102	BIC:	-1241.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.0253	0.002	13.793	0.000	0.022	0.029
x2	0.0157	0.002	6.553	0.000	0.011	0.020
x3	0.0236	0.003	8.150	0.000	0.018	0.029
x4	-0.0038	0.003	-1.216	0.224	-0.010	0.002
x5	0.0078	0.003	2.383	0.017	0.001	0.014
x6	-0.0124	0.004	-3.540	0.000	-0.019	-0.006
x7	0.0060	0.004	1.398	0.162	-0.002	0.014
x8	0.0404	0.005	7.936	0.000	0.030	0.050

Omnibus:	3867.444	Durbin-Watson:	0.142
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49965.579
Skew:	3.722	Prob(JB):	0.00
Kurtosis:	16.389	Cond. No.	2.77

Figure 5: OLS regression results on dimensionality reduced data

- To combat the problem of extremely few data points of patients who have had stroke, I artificially balanced the dataset using SMOTE to a 2:1 ratio of non stroke and stroke patients in order to better train the models to identify any patients who can be at risk of getting a stroke. (figure 6)

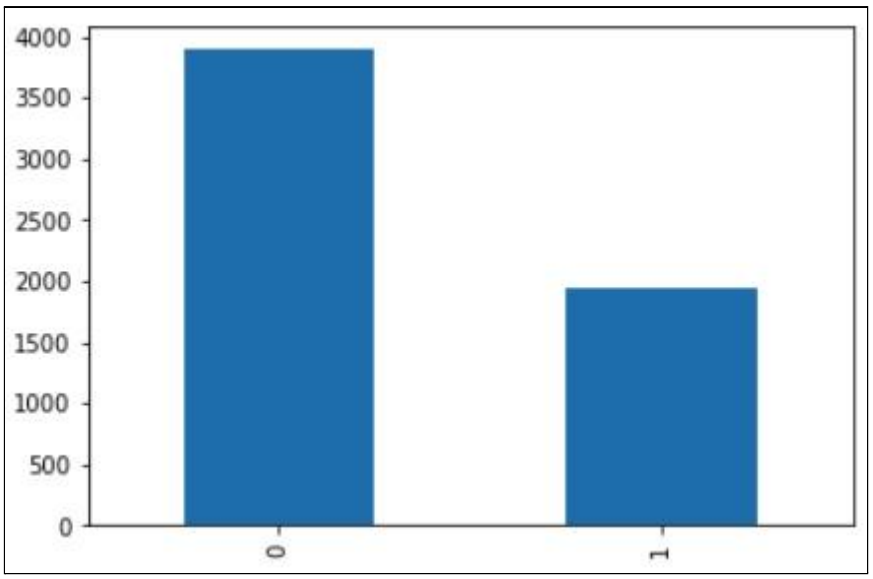


Figure 6: distribution of the stroke variable in the processed data after it has been artificially balanced (0 indicates no stroke, 1 indicates stroke)

Model training: Logistic Regression

- Out of the three different logistic regression models I trained, model 1 had the highest accuracy after 10-fold cross validation. The model was the sklearn logistic regression model with l2 regularization, max iterations of 200 and 'sag' solver. The model had an accuracy of 84.4%, precision of 0.186, recall of 0.574 and f1 score of 0.28 (figure 7)

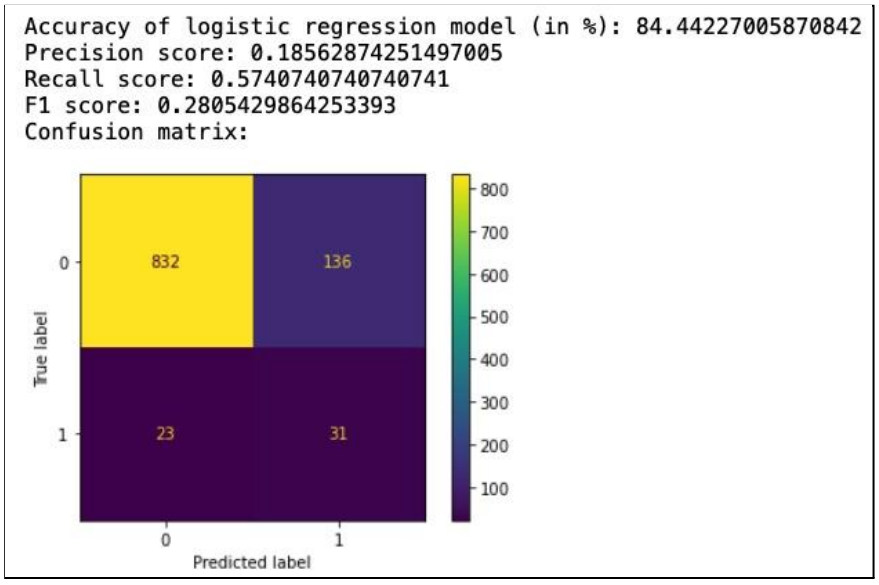


Figure 7: statistics of best logistic regression model after cross validation

- To improve the recall of my model, I reduced the threshold for the same logistic regression model to 0.4 and got better recall but with slightly lower accuracy (figure 8). The number of true positives are higher for this model.

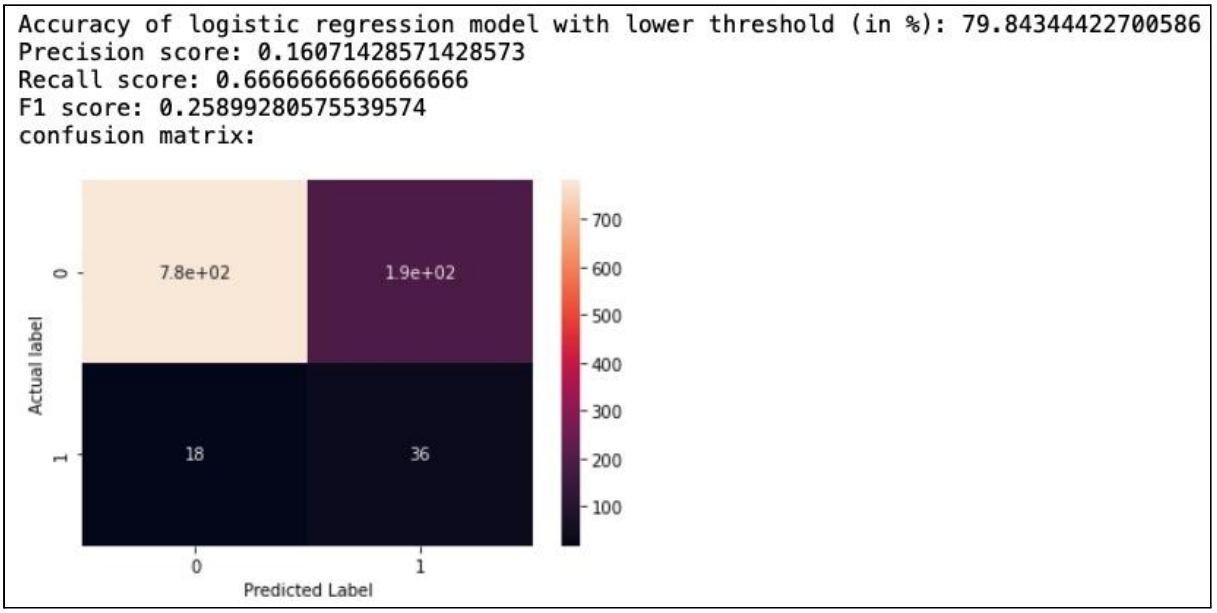


Figure 8: statistics of logistic regression model with lower threshold for returning yes (1) for stroke

Model training: Random Forest Classifier (ensemble)

- I trained 4 different random forest classifiers with different numbers of estimators. Upon running 10-fold cross validation I found that the random forest classifier with bootstrapping set to true, estimators set to 200 and max features set to sqrt got the highest accuracy. The model had a high accuracy of 90.2% on the test data however it had extremely low recall of 0.129 (figure 9).

```
Accuracy of random forrest classifier model (in %): 90.21526418786692
Precision score: 0.11666666666666667
Recall score: 0.12962962962962962
F1 score: 0.12280701754385966
Confusion matrix:
```

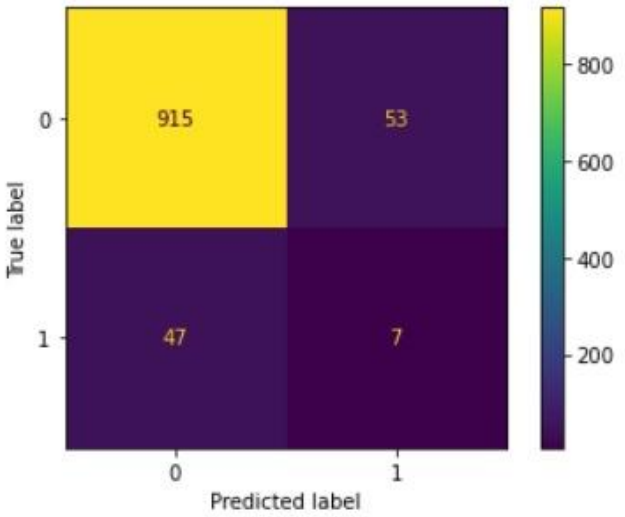


Figure 9: statistics of the best random forrest model

Model training: Neural Networks

- Using keras, I created 3 different neural network architectures. The neural networks varied in the activation functions of their layers and the number of neurons in each layer. After 10-fold cross validation, the best neural network was the network with the following architecture (figure 10). I trained the network for 75 epochs and used a batch size of 10.

```
model_3 = Sequential()
model_3.add(Dense(12, input_dim=8, activation='relu'))
model_3.add(Dense(20, activation='relu'))
model_3.add(Dense(1, activation='sigmoid'))
model_3.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Figure 10: architecture of my best neural network

- The neural network had a test accuracy of 85.42% (figure 11).

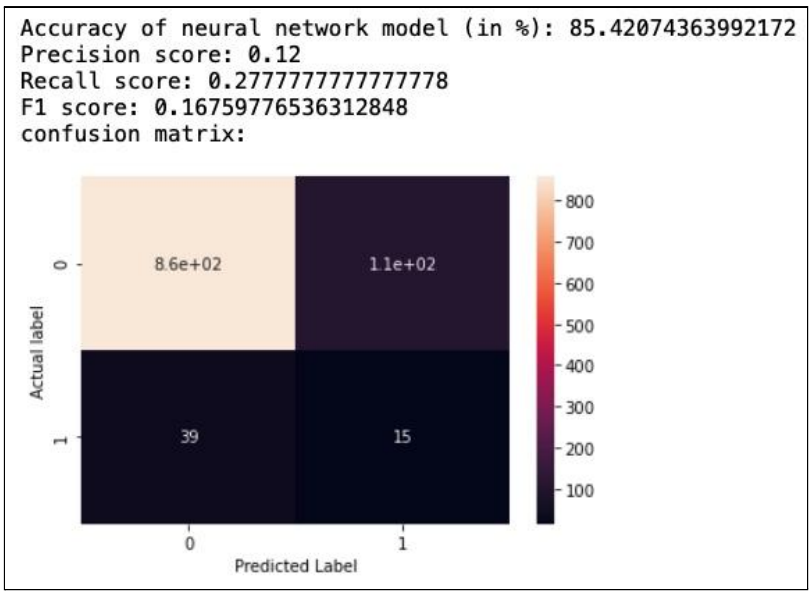


Figure 10: statistics of the best neural network

Discussion

We noticed that some of the features most highly correlated with a person's stroke history are the person's age, their average glucose level and their heart disease history. The CDC has conducted studies and published that the older you get, the more likely you are to get a stroke. This is because your blood vessels and circulatory system in general grows weaker with time increasing the likelihood of a stroke. High glucose levels have also been found to lead to strokes.

We also observed that the type of work doesn't affect a person's chances of getting a stroke and this is true but a high amount of working hours can increase the chance of stroke.

The logistic regression model with the threshold of 0.4 (figure 8) is my recommended model for use if doctors wish to identify if a patient has the medical conditions that can lead to a stroke. I would recommend using the model for initial screening of patients to check if they are at risk of getting a stroke followed by additional analysis of the patients' hydration levels by examining their urine and checking for birth defects with their blood levels for further risk analysis. The models can be further improved if more data is used to train them - 5000 data points is usually not enough to train a sophisticated yet accurate model. Doctors should pay close attention to the glucose levels, number of working hours and blood pressure of patients as they are good indicators of the susceptibility of a patient to stroke, while being extra precautious with older people and people with birth defects related to blood vessels.

Conclusion

In summary, this project included conducting preliminary and supplementary research on strokes to better identify the problem we are dealing with and recognizing the factors that could influence stroke likelihood most. I also conducted some data analysis to find correlations in the data, study the distribution of different variables and to fill missing values.

The data was then sent through a pipeline where numerical variables were standardized to prevent some variables from impacting our models more than others. I also added a relevant feature cross (bmi * glucose level). The categorical variables were one hot encoded. I then conducted OLS regression on the processed data to identify features with high p-values. A lot of features in the processed data had high p-values so I created new data from the processed data using principal component analysis to reduce the dimensionality of the data and reduce collinearity and to also lower the p-values of the features.

Finally, the polished data was run through different models (logistic regression, random forests, neural networks) along with cross validation to identify the best hyperparameters for each model. I also analyzed the models using different metrics like accuracy, recall, precision, f1 score and even confusion matrices to identify the best overall model which was the logistic regression model in figure 8 as it had a decent accuracy score as well as a good recall.

Sources

Centers for Disease Control and Prevention. (2021, May 25). *Stroke facts*. Centers for Disease Control and Prevention. https://www.cdc.gov/stroke/facts.htm.

U.S. Department of Health and Human Services. (2019, July 23). *Researchers get a handle on how to control blood sugar after stroke*. National Institutes of Health. https://www.nih.gov/news-events/news-releases/researchers-get-handle-how-control-blood-sugar-after-stroke.

Scipioni, Jade. (2019, July 31). *Working long hours for a decade can INCREASE stroke risk BY 45%-here's what can help*. CNBC. https://www.cnbc.com/2019/07/03/long-work-hours-could-increase-stroke-risk-heres-what-can-help.html.