Stroke Prediction Analysis using Machine Learning

Birmingham Data Camp-Group 4

Hajisaid, A; Osei,J; Smith, E; Tong,M

Contents

Abstract

Machine learning has been widely used for predictions and forecasting analysis in statistics and data science due to its accuracy and easier modeling without further understanding of the relationship between proxies used for measurement. Machine learning has a big advantage
In this project, we utilize various machine learning models based on the different statistical algorithms/classifiers to predict stroke prevalence using data providing both social and medical information. We find that the random forest and decision trees classifier provides the best prediction accuracy while the data re-balancing method can be achieved using Synthetic Minority Over-sampling Technique "SMOTE" which significantly increases the prediction precision.

1 Introduction

Stroke is a leading causes of death and disability in OECD countries in the UK and US. It causes roughly 75% of deaths from cerebrovascular diseases in the UK in 2018. According to the data from Centers for Diseases and Control and Prevention in US in 2020, 1 in 6 deaths from cardiovascular disease was due to stroke; Every 40 seconds, someone in the United States has a stroke. Every 3.5 minutes, someone dies of stroke. Every year, more than 795,000 people in the United States have a stroke. About 610,000 of these are first or new strokes.About 185,000 strokes—nearly 1 in 4—are in people who have had a previous stroke. About 87% of all strokes are ischemic strokes, in which blood flow to the brain is blocked.Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018. This total includes the cost of health care services, medicines to treat stroke, and missed days of work. Stroke reduces mobility in more than half of stroke survivors age 65 and older.

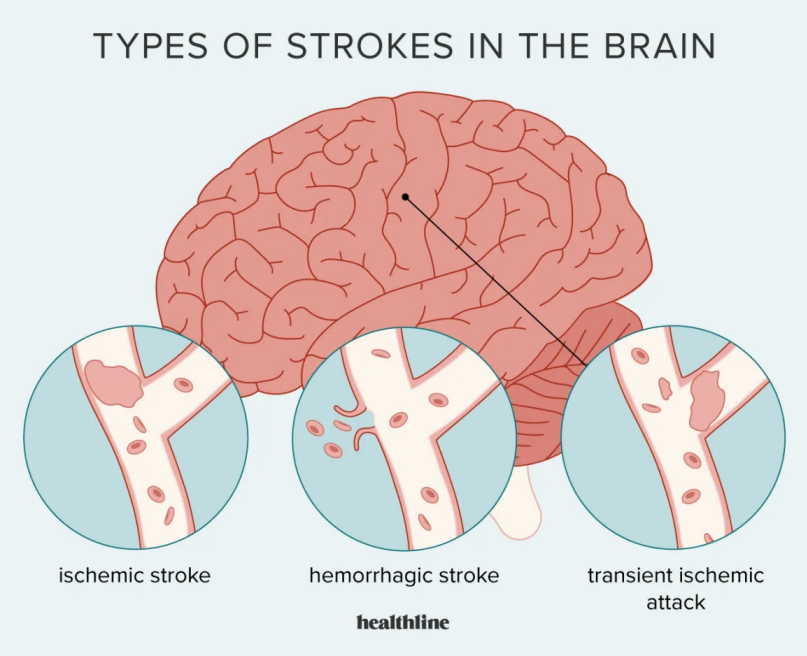

Figure 1: Types of Strokes in the Brain

From the scientific point of view, a stroke is a serious life-threatening medical condition that happens when the blood supply to part of the brain is cut off (NHS). As a cerebrovascular disease, we generally categorize the stroke into three types shown in figure 1. Ischaemic stroke refers the blood supply is stopped because of a blood clot; hemorrhagic stroke refers is defined as a weakened blood vessel supplying the brain bursts; the final type of Stroke is called transient ischaemic attack (TIA), where the blood supply to the brain is temporarily interrupted.

Apart from the direct medical mechanism of stroke prevalence, other factors may also leads to the occurrence of disease. Feigin et al (2016) argue the environmental air pollution as one of the leading risk factors for stroke worldwide. Behavioral risk factors as the majority of risk factors such as smoking, poor diet, and low physical activity counts for 74% of the global burden of stroke. From the 10 leading risk factors for stroke which are high blood pressure, diet low in fruit, high body mass

index (BMI), diet high in sodium, smoking, diet low in vegetables, environmental air pollution, household pollution from solid fuels, diet low in whole grains, and high blood sugar, we could generally categorises those factors into social and medical disease risks factors. Most importantly, the emerging of environmental pollution was found to contribute to one-third (29.2%) of global disability and this is especially high in developing countries (33.7% vs 10.2% in developed countries). To quantify the impacts of environmental air pollution on human health conditions, individuals' health condition measurements(e.g body mass index(BMI)) and living environment (residential region types (urban or county area, etc)) could be good proxies.

In this project, we collect public data including the case with and without stroke embedded with both social and medical information of each individual to predict stroke prevalence. The remaining of the report is organized as follows:in section 2, We introduce how we collect the data and provides statistical analysis after data cleaning; we use various machine learning classifiers and explain the motivations behind. We also discuss the limitations of our results and use data re-sampling method to improve our model and predictions. Finally, we summary and conclude our finding in this project.

2 Data

2.1 Data collection, storage and cleaning

Stroke data we collected for the analysis has been widely used in other projects shared on Kaggle. Our raw data includes 5,110 observations with 12 attributes including numerical values and categorical dummies.Apart from "id", "age" and "gender", attributes in the raw data can be categorised into scientific factors and social factors. The project involved loading the data into S3 buckets on AWS, before using Google Colab to retrieve it. It was imported into SQL to create views for the social and scientific factors of the data, which were then pulled back into Google Colab.

To minimize the impacts of dropping observations to our raw data, we replace all observations that are missing with the mean value of that indicator. For example, if we find one missing value on "BMI", then we use all available values from "BMI" to generate its mean value and replace the missing value of "BMI" with the mean. To mitigate the impacts from outliers, we calculate the statistical prediction value range of each numerical indicators using the formula:

$$Risk\ factors = mean\ value\ of\ risks\ factor \pm 3 * standard\ deviation\ of\ risks\ factor \tag{1}$$

We drop those values fall out of this value range as outliers before running the regression or machine learning training algorithm. We also scale all values to better fits the prediction results of stroke prevalence given by zero and one.

To increase the information set from categorical data, we create dummies variables based on gender, residential types, and working types to group our sample based on their heterogeneity. Among all 5,110 observations, only 249 observations are individuals with stroke showing an highly unbalanced data for our prediction analysis.[1] In model 3, we apply data re-sampling method to overcome this shortage before running the machine learning model.

2.2 Statistical analysis

Table 2 and 3 below shows the statistical summary of all numerical indicators after data cleaning. From the sample summary, we have the hear disease occurrence rate over 5% which is close to

[1]From our sample, we could roughly calculate the stroke prevalence rate is 4.87% by using 249 deviated by 5110.

Table 1: Data Summary

Variable name	Descriptions
Gender (boolean)	Male or female
Age	Numercial value
Hypertension(boolean)	Hypertension
Heart disease(boolean)	Heart disease
Ever married(boolean)	Marriage history
Work type(dummy)	Type of work involved
Residence type(boolean)	Live in county or urban area
Average glucose level	Numerical value
bmi(numerical))	body mass index (BMI)
Smoking status(boolean)	Smoking or not
Stroke(boolean)	Stroke case as patience or not

Table 2: Statistical summary

	age	hypertension	heart disease	average glucose level	bmi	stroke dummy
count	5110	5110	5110	5110	5110	5110
mean	43.2266	0.0975	0.05401	106.1477	28.8932	0.0487
std	22.6126	0.2966	0.2261	45.2836	7.699	0.2153
min	0.08	0.00	0.00	55.12	10.30	0.00
25%	25	0.00	0.00	77.245	23.80	0.00
50%	45.0	0.00	0.00	91.8850	28.40	0.00
75%	61.00	0.00	0.00	114.09	32.80	0.00
max	82.00	1.0	1.0	271.74	97.60	1.00

the probability of stroke prevalence rate statistically. While mean "BMI" value from our sample suggests that our sample individuals are "overweight".[2] The mean value of other indicators reflecting individuals' health condition do not show any selection bias.[3]

[2]The BMI is a convenient rule of thumb used to broadly categorize a person as underweight, normal weight, overweight, or obese based on tissue mass (muscle, fat, and bone) and height. Major adult BMI classifications are underweight (under 18.5 kg/m2), normal weight (18.5 to 24.9), overweight (25 to 29.9), and obese (30 or more) (WHO,2005).

[3]Apart from "BMI" measurement we mentioned in previous, the mean value of hypertension, heart disease, average glucose level all fall into the normal value range.

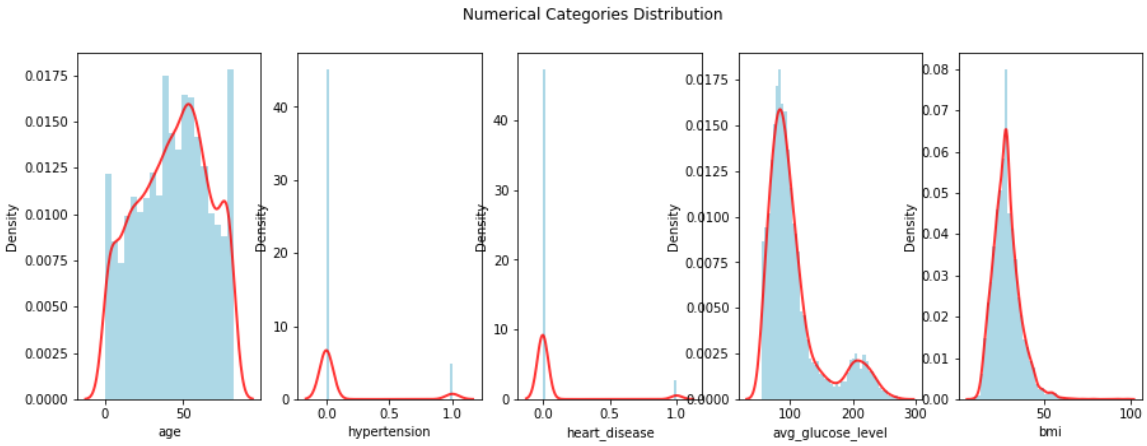

Figure 2: Statistical summary

2.3 The determines of stroke prevalence

In figure 3, we further compare the sample value distribution for each indicator. We still cannot find any significant sample bias between group with stroke and without stroke. For example, sample shares with marriage history or not that have been selected into our data based on stroke dummy is consistent.

On the other hand, we find age and "bmi" are more likely to be the determined indicator of stroke. In figure 4 and 5, we find that stroke are more likely to happen on older individual and those with higher blood pressure. The stroke prevalence (in red) are much higher than those without stroke (in green) when "BMI" value approximately equal to 28 suggesting that "overweight" are more likely leads to stroke.

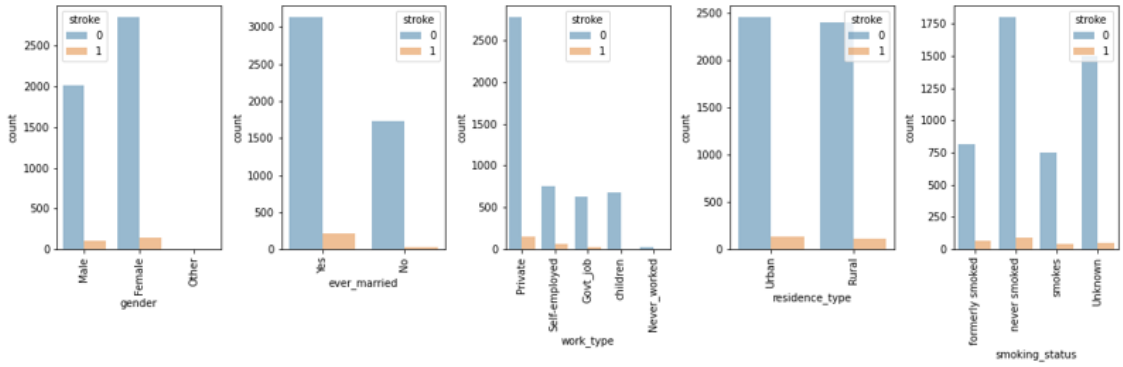

Figure 3: Statistical summary

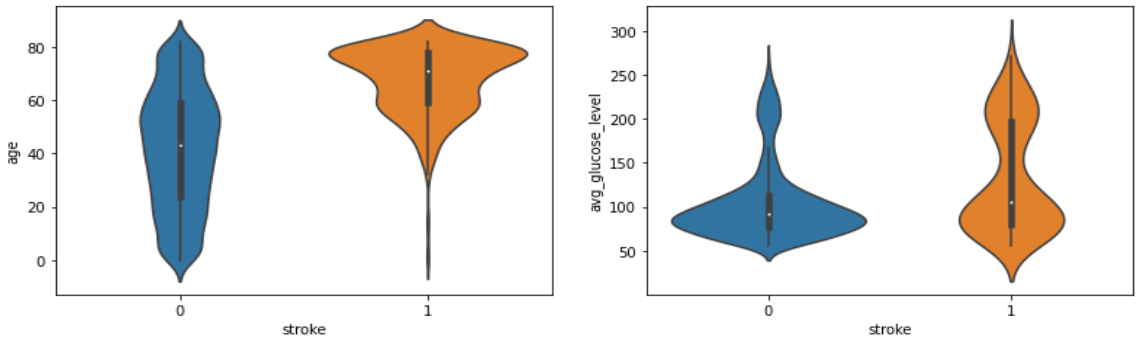

Figure 4: Statistical analysis: Stroke distribution by indicator

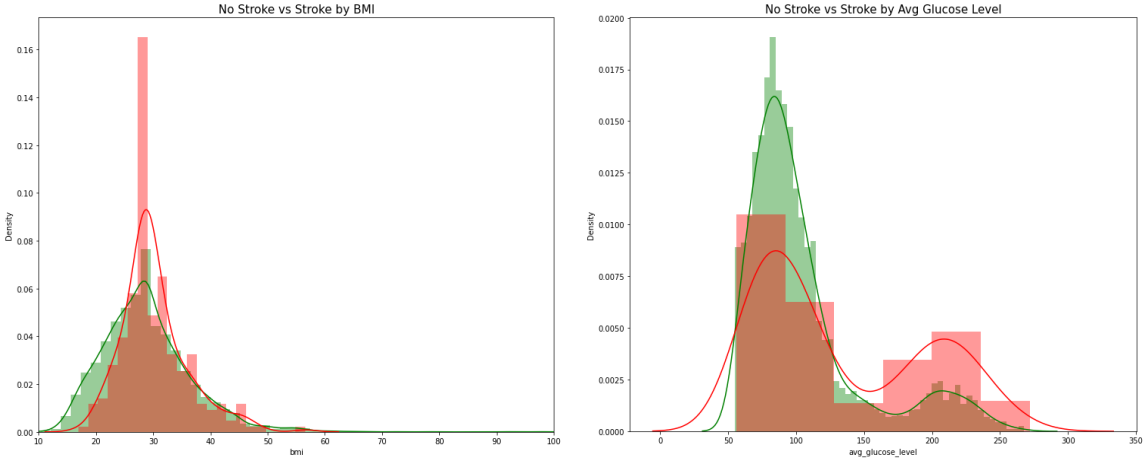

Figure 5: Statistical analysis: Stroke distribution by indicator

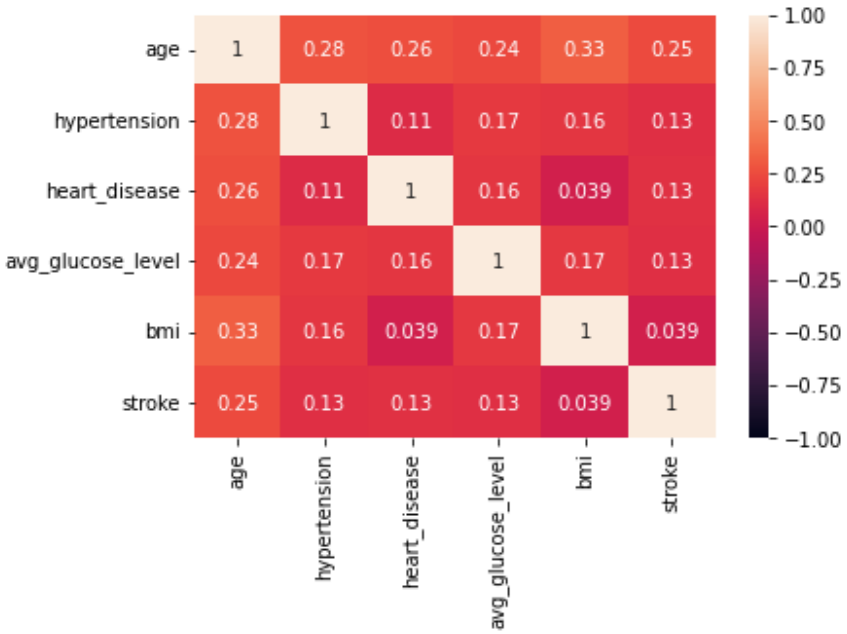

Figure 6: Statistical summary

Finally, to ensure there is no multicollinearity issue if we include all our indicators into the econometrics regressions to compare with machine learning approach, we use Pearson correlations to examine if any pairs of indicators have very high correlations. From figure 6, we find all the correlations are less than 0.4 suggesting each pair of variables is statistical different. Therefore, we are able to include all variables to construct our quantitative analysis using different machine learning classifiers and compare the results with econometrics analysis.

3 Quantitative analysis

3.1 Model construction

Our quantitative modelling follows three steps based on different strategic assumptions and motivations. First, by comparing the prediction results with the econometric regression results, we aim to confirm that machine learning is a good approach to achieve a prediction analysis with high accuracy. Then we try to prove the statistical significance of our prediction model and results. Finally, we aim to improve our model by applying various data processing methods or machine learning algorithm based on different assumptions.

To obtain the best prediction accuracy among all models, we use mean value of cross-validation to measure the prediction accuracy of each classifier provides and activating the machine learning model after adjusting the hyper-parameters of an estimator using "GridSearchCV" to obtain the best cross-validation score.

Table 3 shows the mean value of predictions using different classifiers. Since logistic regression and linear support vector as the linear econometric regressions have better performance than machine learning models utilized, we aim to use the only machine learning classifier with prediction accuracy close to the best prediction accuracy given by regressions to complete the machine learning algorithm

Table 3: Testing prediction accuracy by different models

Cross validation mean values for different classifier	
Type of Classifier	Prediction mean value
Decision Tree	0.915
Logistic Regression	0.952
Linear Support Vector (SVC)	0.952
AdaBoost	0.917
Random Forest	0.951
Gradient Boosting	0.9506
K-nearest Neighbors (k-NN)	0.9488

predictions. Figure 7 shows the prediction accuracy given by cross validation score is close to the prediction score using training data approximately equal to 95%. Recall our sample with stroke also counts for about 95% of the full sample, it is reasonable to believe the random forest may obtain at least the same prediction accuracy compared with other statistical analyses and method used.

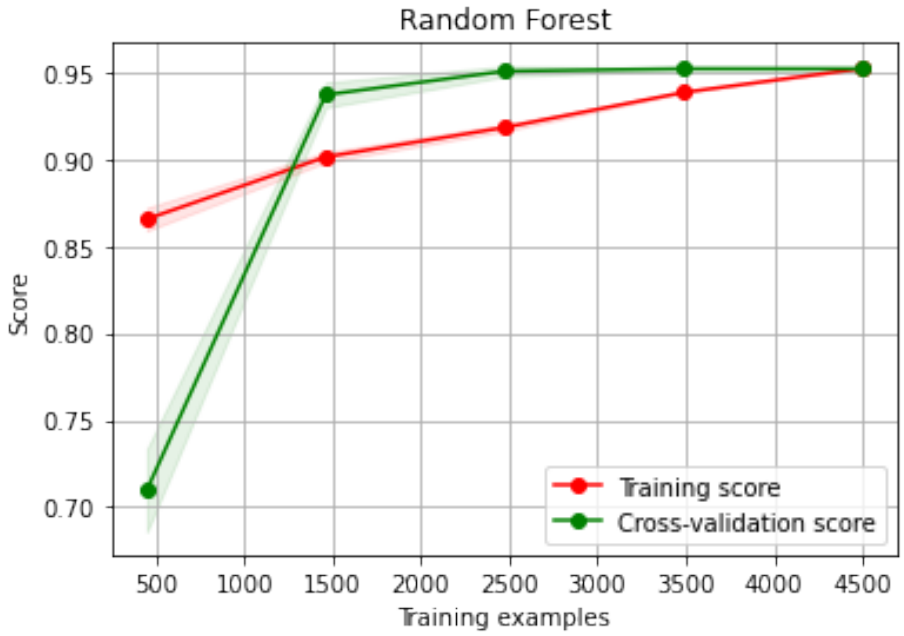

Figure 7: Prediction results by random forest

However, the analysis overlook other indicators to examine the significant results using this prediction model. Precision and confusion matrix has not been estimated and considered. Therefore, we further explore the supervised machine learning model package used in the next step.

3.2 Failure of significant predictions

In the second modelling strategies, we apply more machine learning classifiers to construct prediction models to compare with linear regressions. To solve the unbalanced data issue, we apply

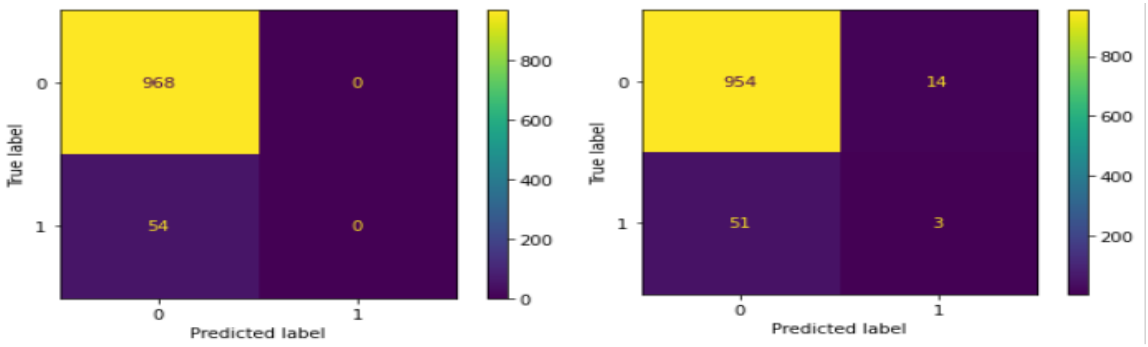

Figure 8: Confusion matrix Logistic regression vs. KNeighbors

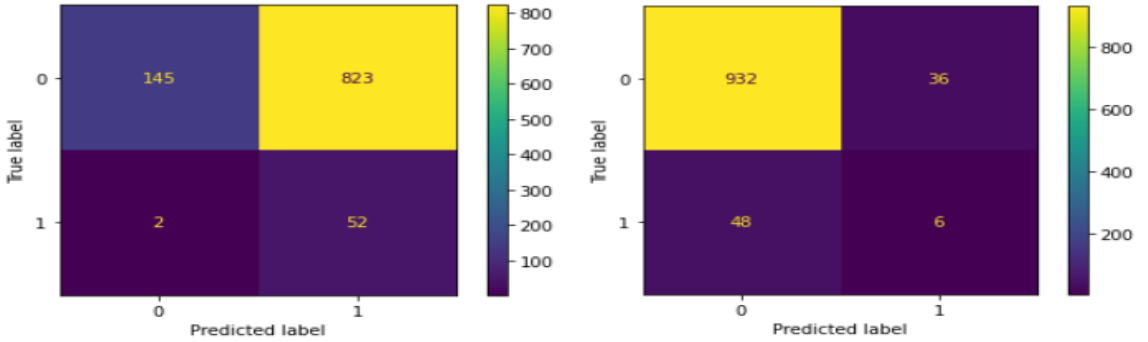

Figure 9: Confusion matrix Gaussian Naive Bayes (GaussianNB) vs. Bernoulli Naive Bayes

Synthetic Minority Oversampling Technique (SMOTE) to re-sampling training data help to prove the statistical significance of our predictions.

After scaling and splitting into training and testing data, we had our prediction results in Confusion matrix in Figure 8-12 and Table 4. Although the confusion matrix results show a good performance that the dominated share of predictions falls into predicting truth/false correctly, the precision of all models are quite poor. F1 score for all prediction models are less than 20% while a reasonable value range to define a prediction model with significant accuracy is over 80%. This can be due to the unbalanced data caused selection bias issue since we only re-sample the training data. We improve the precision in the next step of the modeling construction.

On the other hand, prediction accuracy scores are still quite high suggesting the capability of predicting values from models apply is with good performance. We select the random forest as the machine learning algorithm with highest prediction accuracy altogether with two linear regression prediction classifiers with highest prediction accuracy (logistic regression and SVC) as our target model to further improve their prediction accuracy by selection best parameters. Using "GridSearchCV" from "sklearn" package in python, we loop through predefined hyper-parameters and fit your estimator (model) on your training set. Table 5 summaries the best parameters for each model that results in the best prediction accuracy.

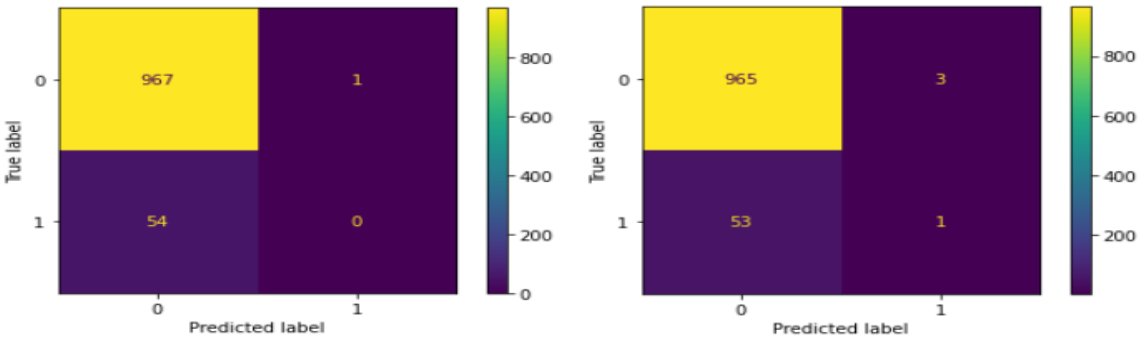

Figure 10: Confusion matrix Decision Tree vs. Random Forest

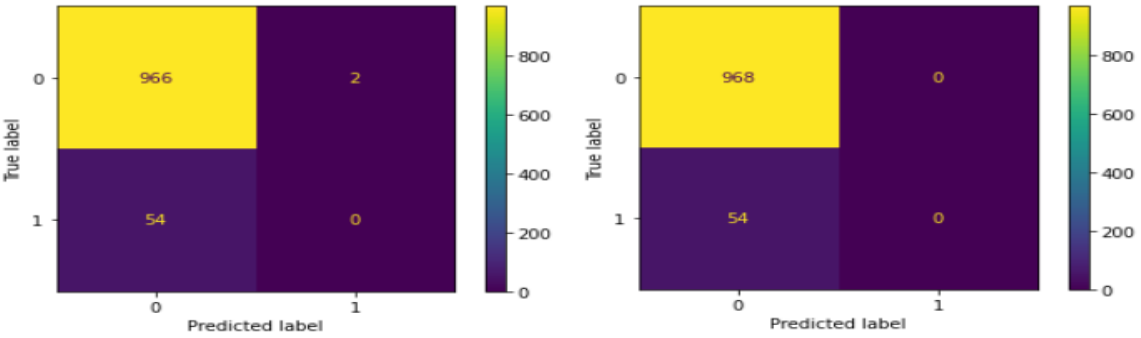

Figure 11: Confusion matrix Extreme Gradient Boosting(XGBoost) vs. Linear Support Vector Classification(SVC)

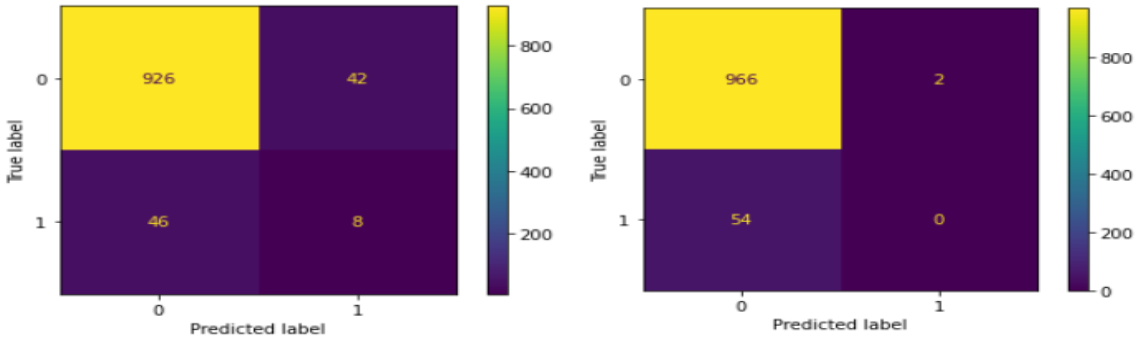

Figure 12: Confusion matrix Extremely Randomized Trees vs. AdaBoost

Table 4: Prediction estimation summary

Prediction estimation summary by different classifier					
Name of classifier	Accuracy Score	K Fold Accuracy	F1 Score	Recall	ROC
Logistic Regression	0.947	95.18	0.0	0.0	0.5
KNeighborsClassifier	0.936	93.91	0.0845	0.0556	0.5205
GaussianNB	0.193	18.57	0.1119	0.963	0.556
BernoulliNB	0.918	92.59	0.125	0.111	0.537
Decision Tree Classifier	0.946	95.18	0.0	0.0	0.499
Random Forest Classifier	0.945	95.11	0.034	0.0185	0.5077
XGBClassifier	0.945	95.23	0.0	0.0	0.499
SVC	0.947	95.23	0.0	0.0	0.5
Extra Tree Classifier	0.914	91.83	0.154	0.148	0.552
Ada Boost Classifier	0.945	95.23	0.0	0.0	0.499

Table 5: Statistical summary

Best parameters summary		
Name of classifier	Best Accuracy	Best Parameters
Logistic Regression	95.23%	'cv': 5, 'random state': 0
Random Forest Classifier	95.03%	'number of estimators': 75
SVC	95.23%	'degree': 2

3.3 Data re-sampling using SMOTE

The lesson from the second modeling strategy tells us the importance of data processing. In the final step, we re-sampling all observations (including both training and testing data) using SMOTE and proceed the similar machine learning predictions. From results in Figure 13 and 14 and Table 6, the prediction accuracy for logistic regression is still over 90% while other machine learning models experience a significant improvement after data processing of full sample. More importantly, precision value also show a high significant of prediction with f1-score over 90%.

Regarding model comparisons, we find random forest and extra tree (decision tree) have the same and best prediction results with both accuracy and precision equals 97%. We may conclude that the prediction results using machine learning exceed the statistical mean value we generated in the first step of modeling construction.

4 Conclusions

In this project, we use various machine learning algorithm in combine with various data processing assumptions/strategies to predict stroke prevalence using public data collected from kaggle website. After data cleaning and applying the data re-sampling process, the prediction results using machine learning is better than econometrics regressions. Random forest and decision tree can be good option help with accurate prediction analysis using medical data although some data processing and presumptions should be made.

Some explorations of further improve this analysis to collect more comprehensive micro-level information to help with different presumption modelings. Since the emerging of air pollution caused risks to stroke prevalence attracts more attention, the potential caveats based on the data

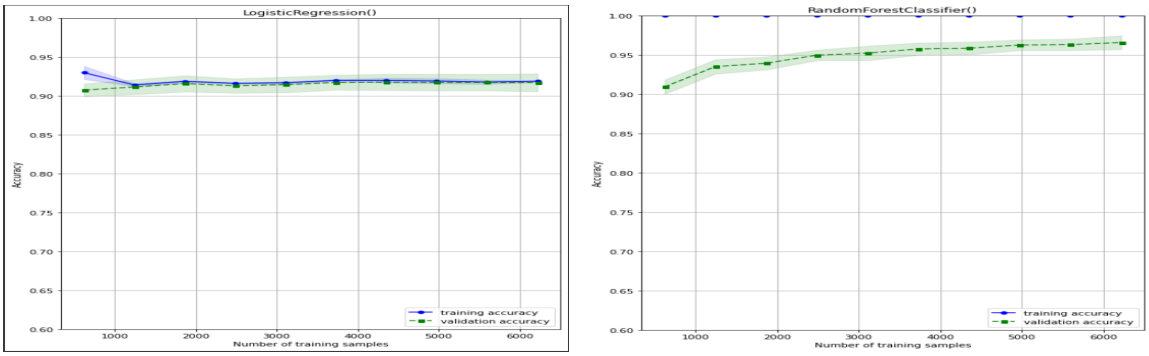

Figure 13: Prediction accuracy for logistic regressions and random forest algorithm ML

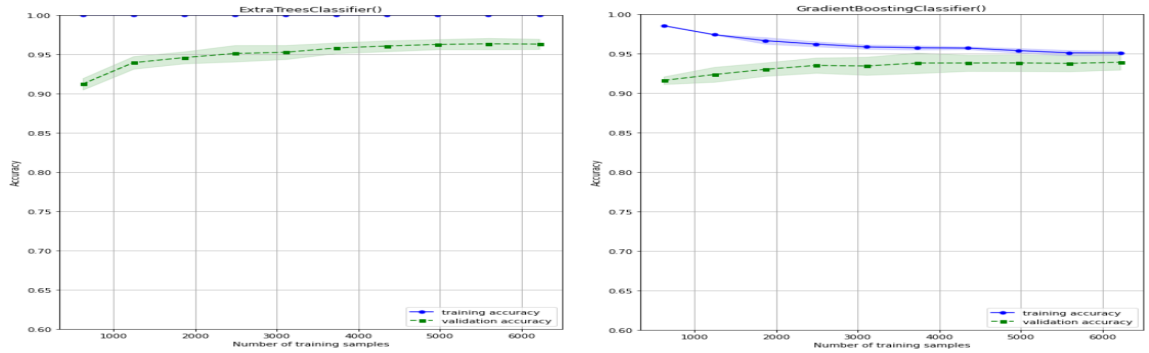

Figure 14: Prediction accuracy for Extra tree and Gradient Boost algorithm ML

Table 6: Prediction results summary for final step modeling constructions

Indicators	Logistic regressions				Random forest			
	precision	recall	f1-score	support	precision	recall	f1-score	support
0	0.89	0.95	0.92	970	0.96	0.98	0.97	970
1	0.95	0.89	0.92	974	0.96	0.98	0.97	970
accuracy			0.92	1944			0.97	1944
macro avg	0.92	0.92	0.92	1944	0.97	0.97	0.97	1944
weighted avg	0.92	0.92	0.92	1944	0.97	0.97	0.97	1944
Indicators	Extra trees				Gradient Boost			
	precision	recall	f1-score	support	precision	recall	f1-score	support
0	0.96	0.98	0.97	970	0.93	0.96	0.95	970
1	0.98	0.96	0.97	974	0.96	0.93	0.94	974
accuracy			0.97	1944			0.95	1944
macro avg	0.97	0.97	0.97	1944	0.95	0.95	0.95	1944
weighted avg	0.97	0.97	0.97	1944	0.95	0.95	0.95	1944

we have is to provide sub-sample analysis by splitting social and scientific variables to examine the importance of stroke prevalence from social or scientific factors.