Stroke Prediction using Machine Learning Models

Project 4

Group 4- Hajisaid, A; Osei,J; Smith, E; Tong,M

7^{th}, May 2022

1. Introductions
   - Motivations
2. Data
   - Data collection
   - Statistical analysis
3. Methodology
   - Model 1
   - Model 3

Motivations

- Stroke is a type of cerebrovascular disease, which is one of the leading causes of death and disability in the UK.
- Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018.
- Stroke is a leading cause of serious long-term disability.

Stroke statistics in the UK

- Stroke accounts for roughly 75% of deaths from cerebrovascular diseases. 100,000 people have strokes each year.

- Stroke prevalence rate 2% in average while prevalence in total is 1,291,890 in the UK.

- The amount of hospital admissions are 136,345 in total.

Stroke statistics in US

- In 2020, 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

Summary

The determines of stroke prevalence

Feigin et al (2016) conclude that the emerging of environmental air pollution become one of the leading risk factors for stroke worldwide

Behavioral risk factors such as smoking, poor diet, and low physical activity are the majority risk factors over 74% counts for the global burden of stroke

It implies social behaviors is also important for stroke prediction and could be the good reason for us to use both medical and social factors to predict stroke prevalence

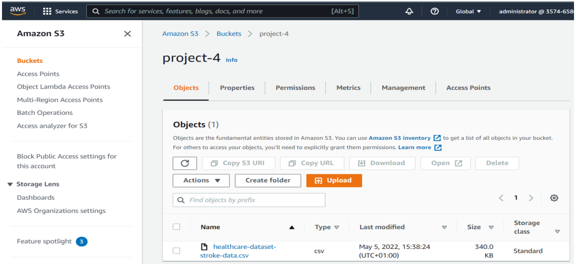
Motivations

In this project, we compare machine learning models using various attributes in related with health conditions(gender, ages, smoking, heart disease, BMI, etc.) helping to predict strokes.

Data collection and processing

- Data we utilize in this project is retrieved from Stroke Prediction Dataset that has been widely used on Kaggle
- The data includes 5,110 observations with 12 attributes capturing the health condition of individuals
- After introducing the raw data, we first find if any missing values for all proxies
- We find 'bmi' has some values that is "N/A", we fill in those values using the mean of "bmi"
- We also drop the "id" column representing insignificant information to support our prediction estimations
- Among 5,110 observations, we have 249 patience cases with stroke on record (Unbalanced data)

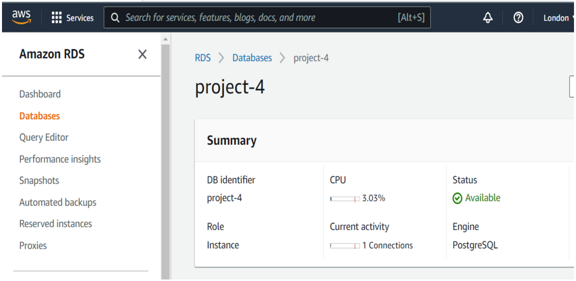
Data storage



Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use **Amazon S3 inventory** to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. **Learn more**

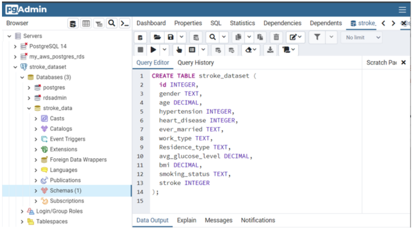
□	Name ▲	Type	Last modified	Size	Storage class
□	healthcare-dataset-stroke-data.csv	csv	May 5, 2022, 15:38:24 (UTC+01:00)	340.0 KB	Standard



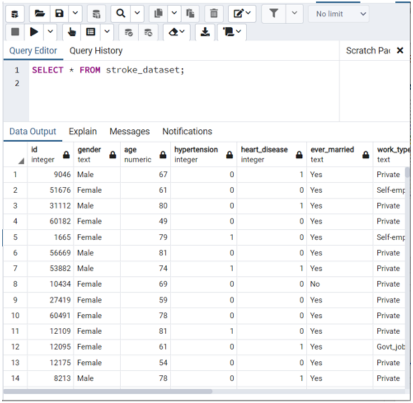
Summary

DB identifier	CPU	Status
project-4	3.03%	Available
Role	Current activity	Engine
Instance	1 Connections	PostgreSQL

Data collection and processing



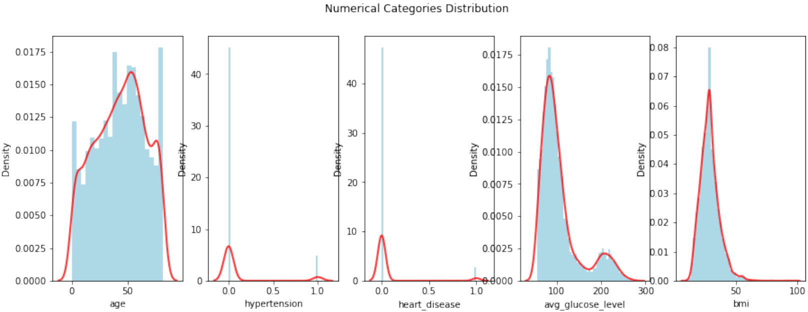
```
CREATE TABLE stroke_dataset (
  id INTEGER,
  gender TEXT,
  age DECIMAL,
  hypertension INTEGER,
  heart_disease INTEGER,
  ever_married TEXT,
  work_type TEXT,
  Residence_type TEXT,
  avg_glucose_level DECIMAL,
  bmi DECIMAL,
  smoking_status TEXT,
  stroke INTEGER
);
```



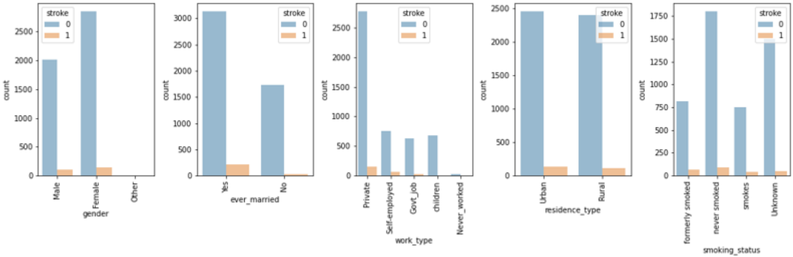
```
SELECT * FROM stroke_dataset;
```

	id integer	gender text	age numeric	hypertension integer	heart_disease integer	ever_married text	work_type text
1	9046	Male	67	0	1	Yes	Private
2	51676	Female	61	0	0	Yes	Self-emp
3	31112	Male	80	0	1	Yes	Private
4	60182	Female	49	0	0	Yes	Private
5	1665	Female	79	1	0	Yes	Self-emp
6	56669	Male	81	0	0	Yes	Private
7	53882	Male	74	1	1	Yes	Private
8	10434	Female	69	0	0	No	Private
9	27419	Female	59	0	0	Yes	Private
10	60491	Female	78	0	0	Yes	Private
11	12109	Female	81	1	0	Yes	Private
12	12095	Female	61	0	1	Yes	Govt_job
13	12175	Female	54	0	0	Yes	Private
14	8213	Male	78	0	1	Yes	Private

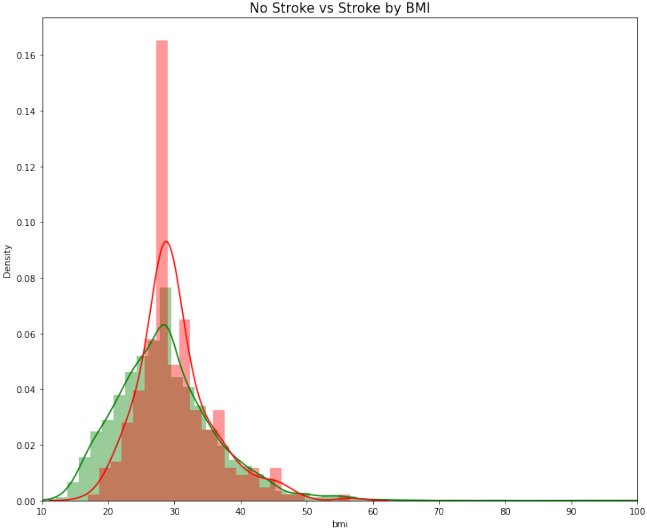
Statistical analysis



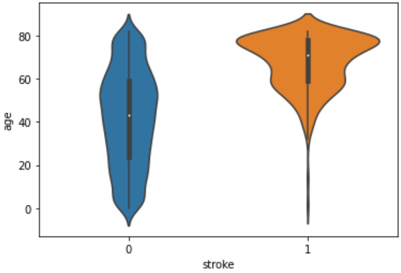
Data statistical analysis

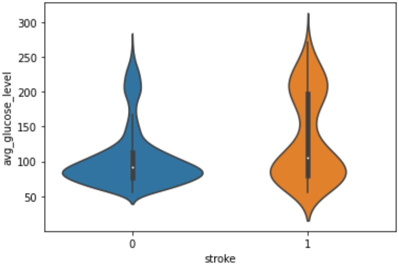


Data statistical analysis

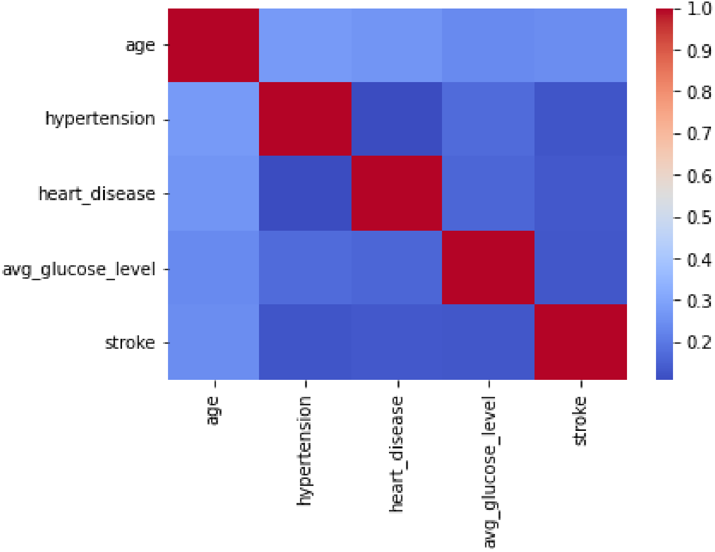


Data statistical analysis





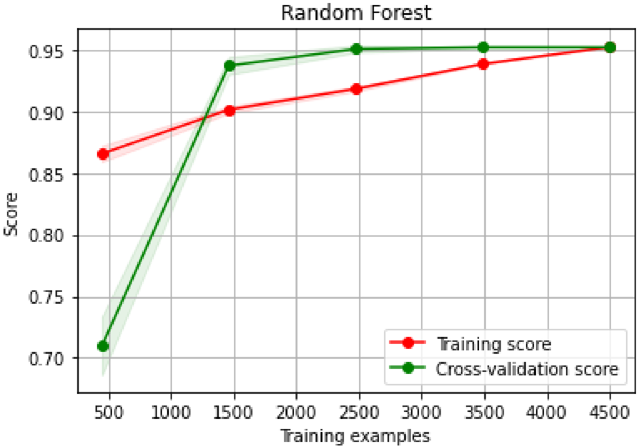
Statistical analysis



Machine learning model comparisons

- We standardize the data using statistical formula before construct the prediction model for econometrics regressions and machine learning predictions.
- We try three-round of the prediction analysis based on different strategies.
- The prediction accuracy are mainly considered in model 1 and 2 while prediction accuracy and precision are both considered in model 3.
- In the first round analysis, we generate the classifier based on different prediction algorithm aim to compare their cross-validation mean prediction accuracy using training data
- Decision Tree (0.915), Logistic Regression (0.952), C-support vector (SVC)(0.952) ,AdaBoost (0.915), Random Forest(0.951), Gradient Boosting(0.9506) and KNeighbors(KNNs) (0.9488).

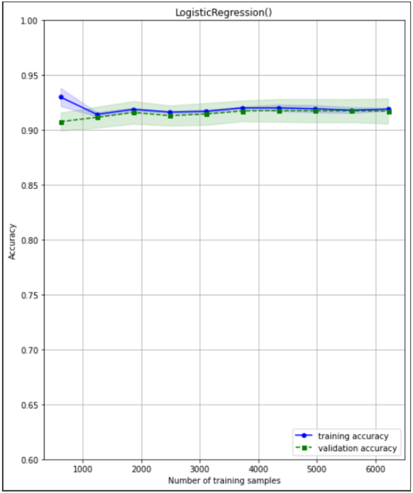
Machine learning model comparisons

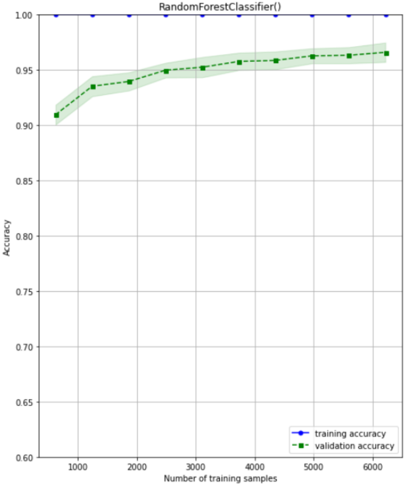


Summary

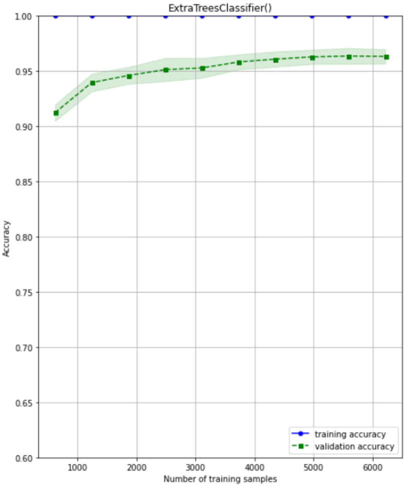
- Random forest is the machine learning model with highest prediction accuracy
- We use Exhaustive Grid Search (GridSearchCV) to adjust the hyper-parameters estimator
- We find a over 95% of the prediction accuracy
- No further analysis on precision and confusion matrix (failure of model 2)
- Synthetic Minority Over-sampling Technique (SMOTE) could be a good data re-sampling method to overcome imbalanced data issue

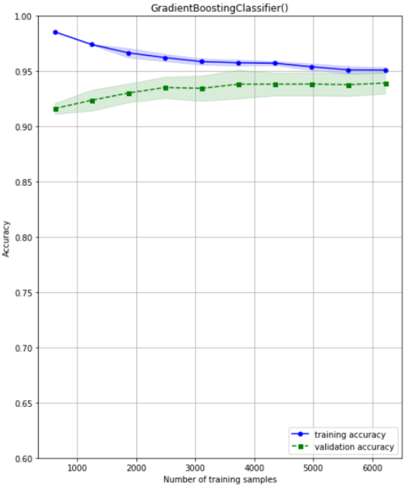
Machine learning model comparisons





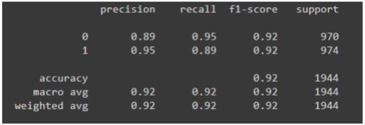
Machine learning model comparisons





Machine learning model comparisons

Logic regression prediction results

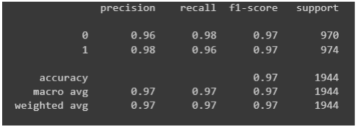


```
              precision    recall  f1-score   support

           0       0.89      0.95      0.92       970
           1       0.95      0.89      0.92       974

    accuracy                           0.92      1944
   macro avg       0.92      0.92      0.92      1944
weighted avg       0.92      0.92      0.92      1944
```

Random Forest prediction results

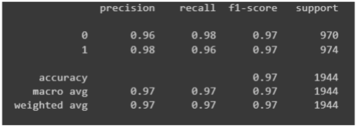


```
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       970
           1       0.98      0.96      0.97       974

    accuracy                           0.97      1944
   macro avg       0.97      0.97      0.97      1944
weighted avg       0.97      0.97      0.97      1944
```

Extra Trees prediction results

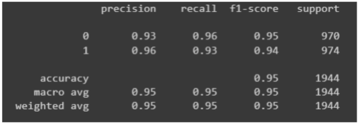


```
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       970
           1       0.98      0.96      0.97       974

    accuracy                           0.97      1944
   macro avg       0.97      0.97      0.97      1944
weighted avg       0.97      0.97      0.97      1944
```

Gradient Boosting prediction results



```
              precision    recall  f1-score   support

           0       0.93      0.96      0.95       970
           1       0.96      0.93      0.94       974

    accuracy                           0.95      1944
   macro avg       0.95      0.95      0.95      1944
weighted avg       0.95      0.95      0.95      1944
```

Conclusions

- Random forest and Extra trees are two classifiers result highest prediction accuracy and precision (F1 score) among four most power method
- "SMOTE" as good re-sampling method may significantly solve imbalanced data issue
- Other Neural network approach could also be applied in the future
- Splitting scientific and social indicators to construct new prediction model for comparisons could be anther good option