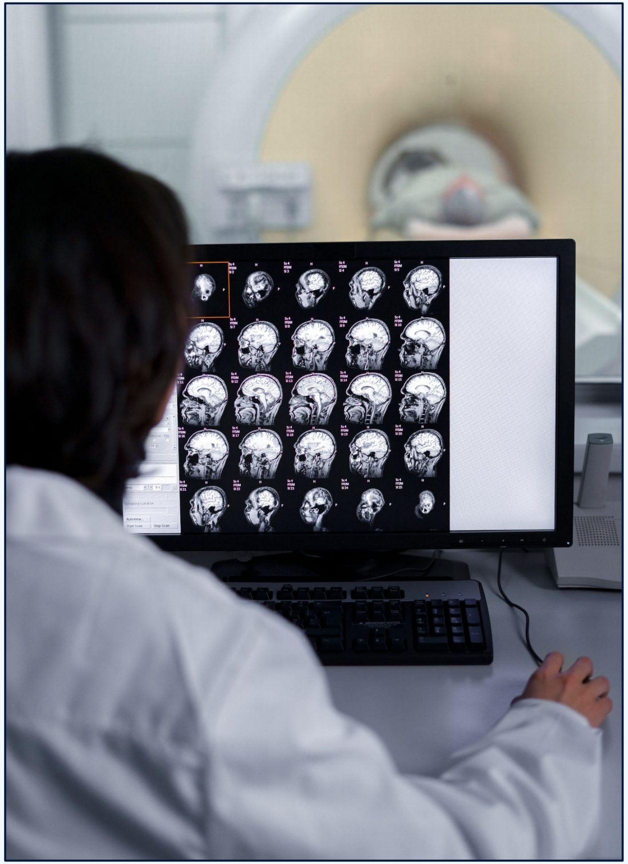


01

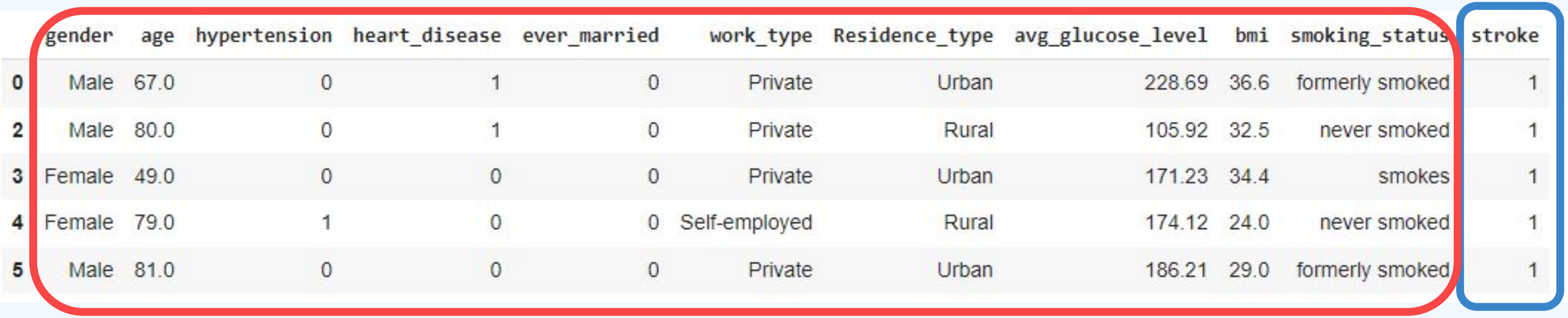
Data preprocess

Dataset Introduction

- **Predictive model**

Feature variables

Target



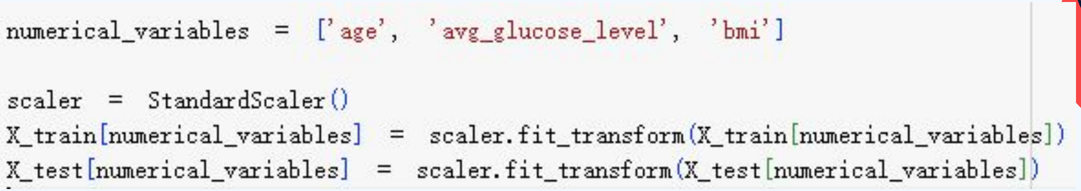
	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	0	Private	Urban	228.69	36.6	formerly smoked	1
2	Male	80.0	0	1	0	Private	Rural	105.92	32.5	never smoked	1
3	Female	49.0	0	0	0	Private	Urban	171.23	34.4	smokes	1
4	Female	79.0	1	0	0	Self-employed	Rural	174.12	24.0	never smoked	1
5	Male	81.0	0	0	0	Private	Urban	186.21	29.0	formerly smoked	1

- **Anomaly Detection**

Data preprocess

Numerical variables

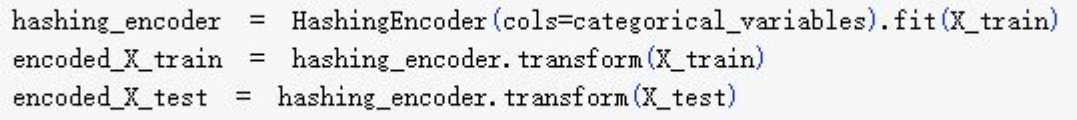
- Standardization



```
numerical_variables = ['age', 'avg_glucose_level', 'bmi']

scaler = StandardScaler()
X_train[numerical_variables] = scaler.fit_transform(X_train[numerical_variables])
X_test[numerical_variables] = scaler.fit_transform(X_test[numerical_variables])
```

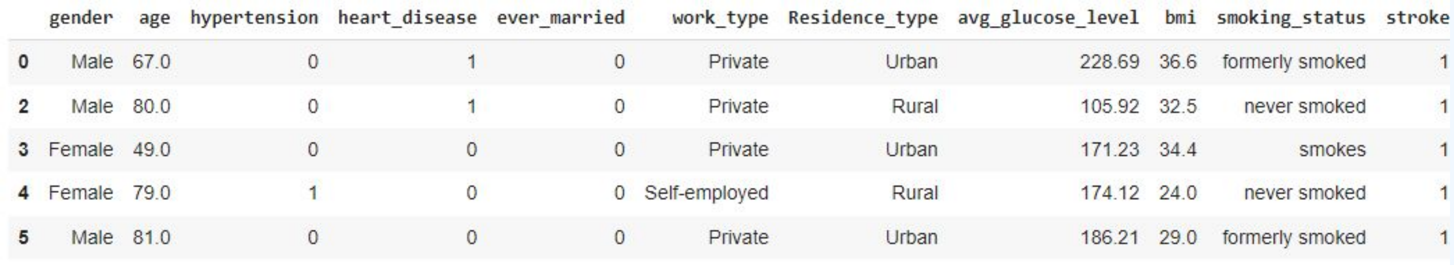
Categorical variables

- Hash encoding

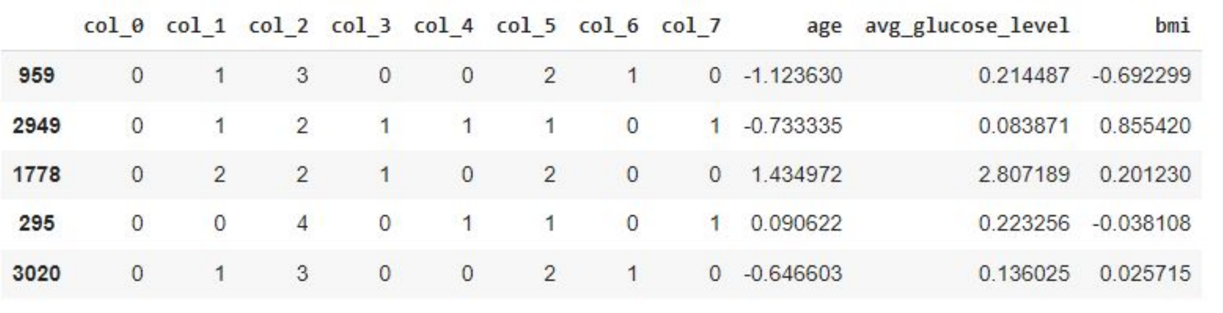


```
hashing_encoder = HashingEncoder(cols=categorical_variables).fit(X_train)
encoded_X_train = hashing_encoder.transform(X_train)
encoded_X_test = hashing_encoder.transform(X_test)
```

After Preprocessing



	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	0	Private	Urban	228.69	36.6	formerly smoked	1
2	Male	80.0	0	1	0	Private	Rural	105.92	32.5	never smoked	1
3	Female	49.0	0	0	0	Private	Urban	171.23	34.4	smokes	1
4	Female	79.0	1	0	0	Self-employed	Rural	174.12	24.0	never smoked	1
5	Male	81.0	0	0	0	Private	Urban	186.21	29.0	formerly smoked	1



	col_0	col_1	col_2	col_3	col_4	col_5	col_6	col_7	age	avg_glucose_level	bmi
959	0	1	3	0	0	2	1	0	-1.123630	0.214487	-0.692299
2949	0	1	2	1	1	1	0	1	-0.733335	0.083871	0.855420
1778	0	2	2	1	0	2	0	0	1.434972	2.807189	0.201230
295	0	0	4	0	1	1	0	1	0.090622	0.223256	-0.038108
3020	0	1	3	0	0	2	1	0	-0.646603	0.136025	0.025715

02

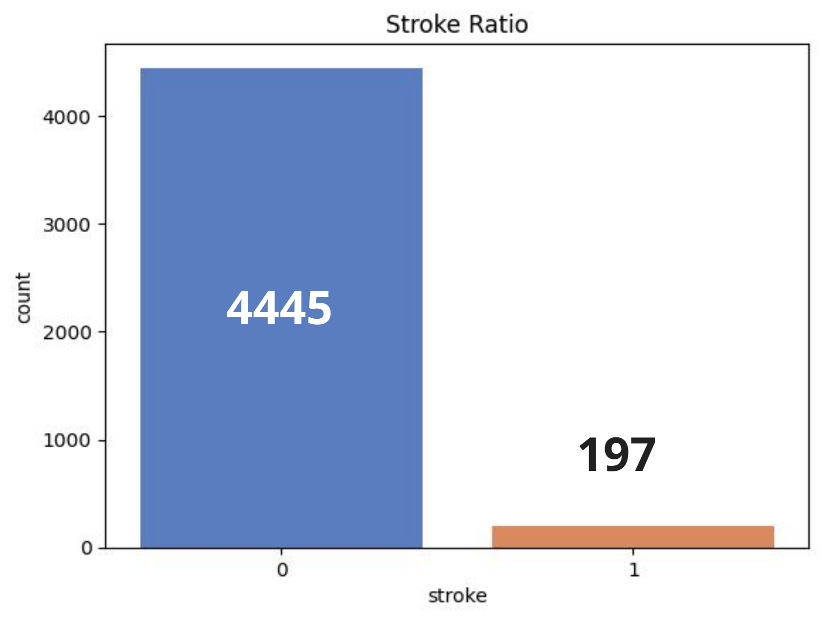
Model Training

Before that

Imbalanced Data

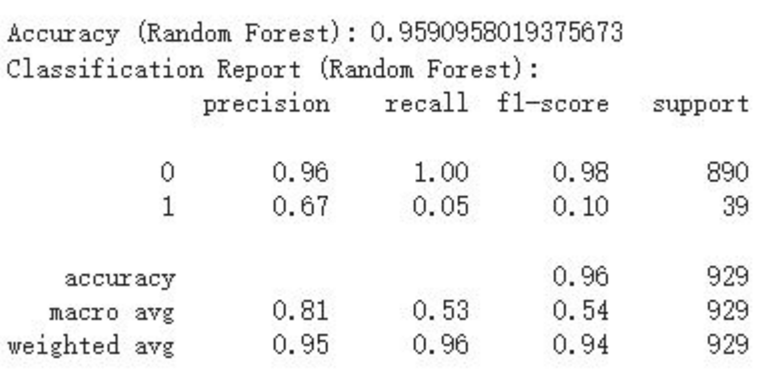
Target Label : **Stroke**

$$\frac{197}{4642} = 0.04$$

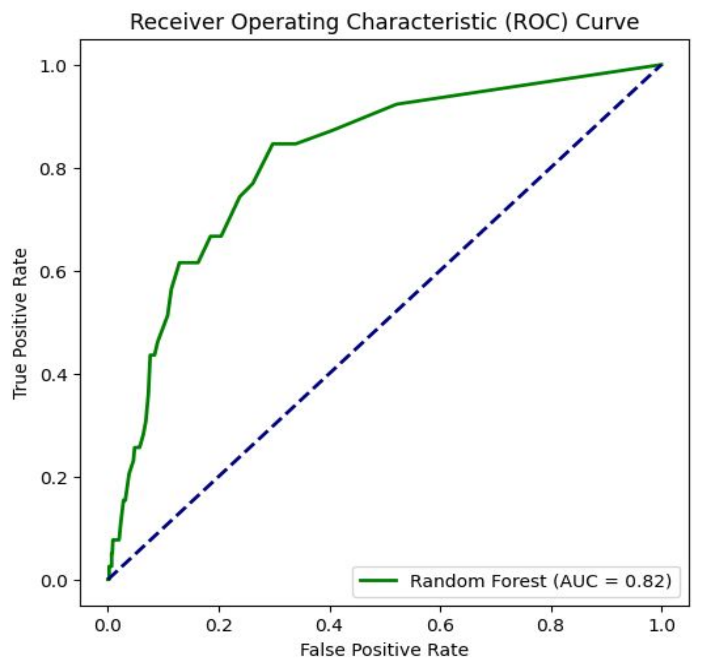


Training (without resampling)

Random Forest

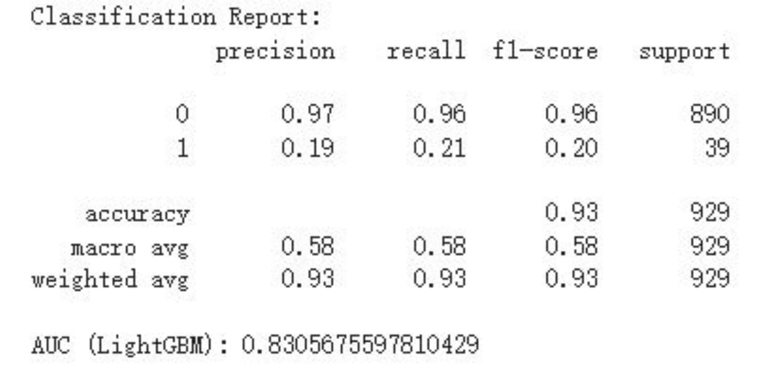


```
Accuracy (Random Forest): 0.9590958019375673
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       0.96      1.00      0.98       890
           1       0.67      0.05      0.10        39

    accuracy                           0.96       929
   macro avg       0.81      0.53      0.54       929
weighted avg       0.95      0.96      0.94       929
```

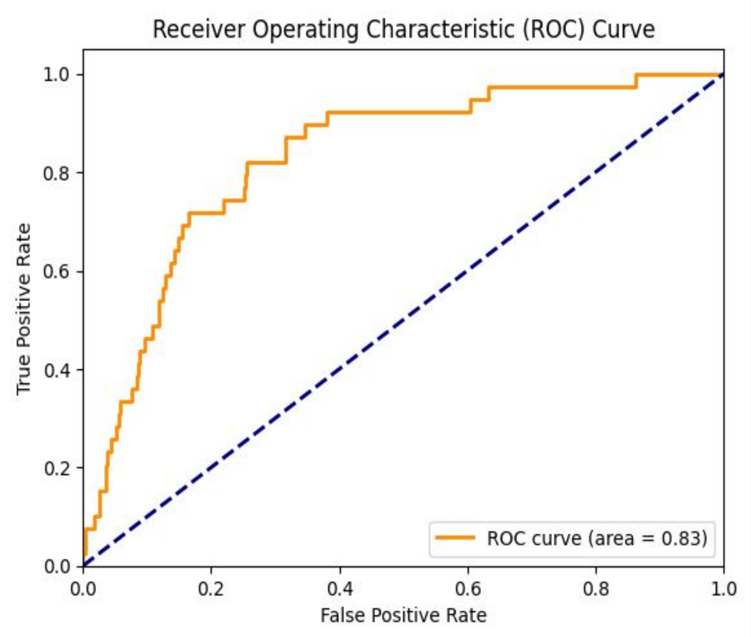


Training (without resampling)

LightGBM



```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.96      0.96       890
           1       0.19      0.21      0.20        39

    accuracy                           0.93       929
   macro avg       0.58      0.58      0.58       929
weighted avg       0.93      0.93      0.93       929

AUC (LightGBM): 0.8305675597810429
```



03

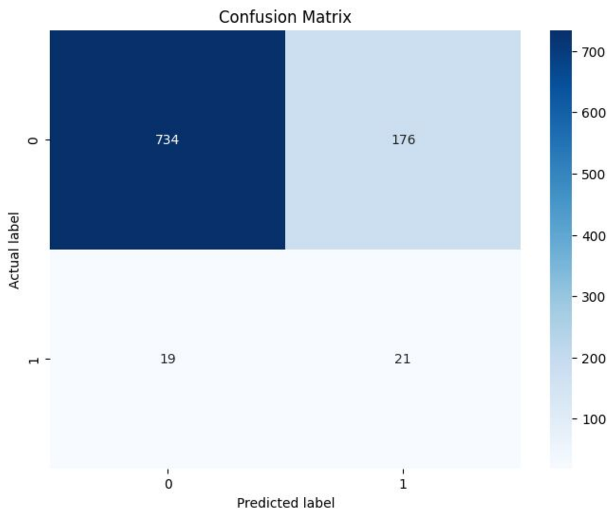
Resampling

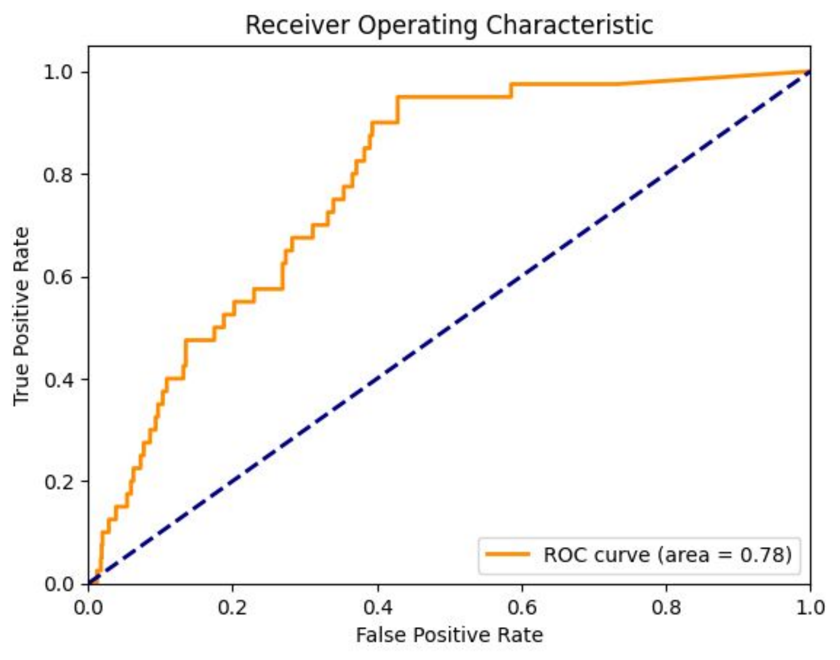
Training (SMOTE + ENN)

Accuracy: 0.79

AUC: 0.66

Random Forest



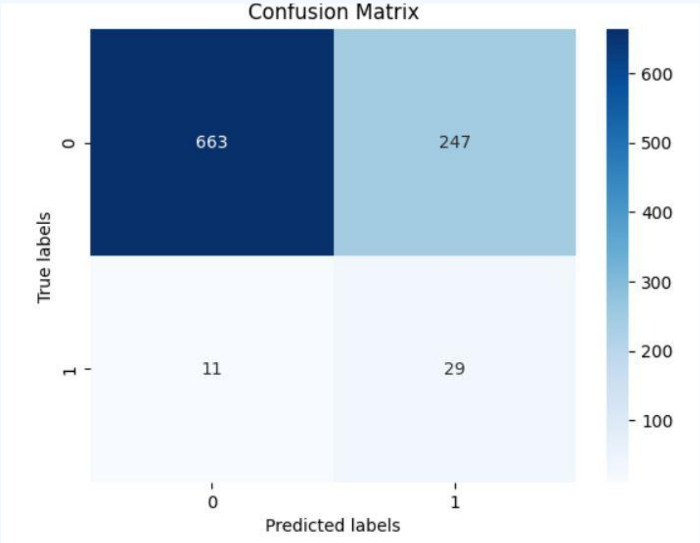


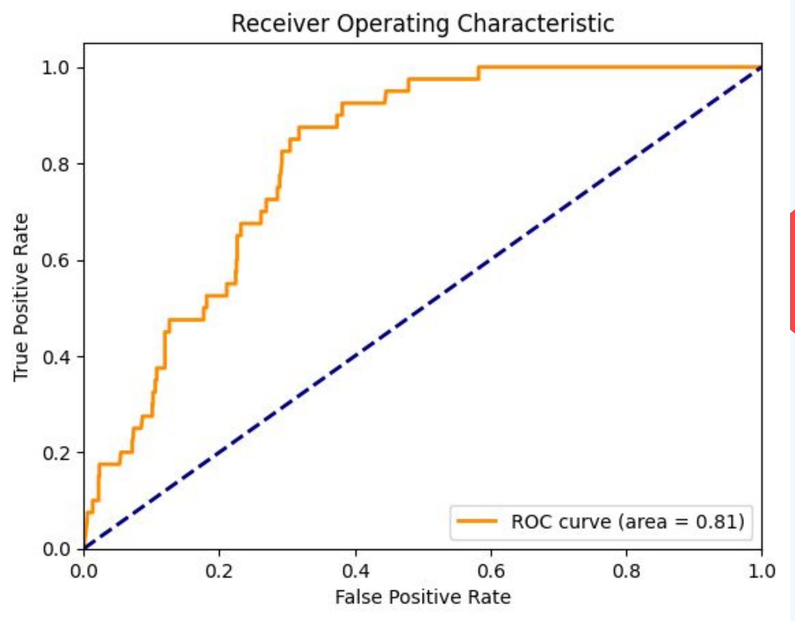
Training (SMOTE + ENN)

Accuracy: 0.72

AUC: 0.81

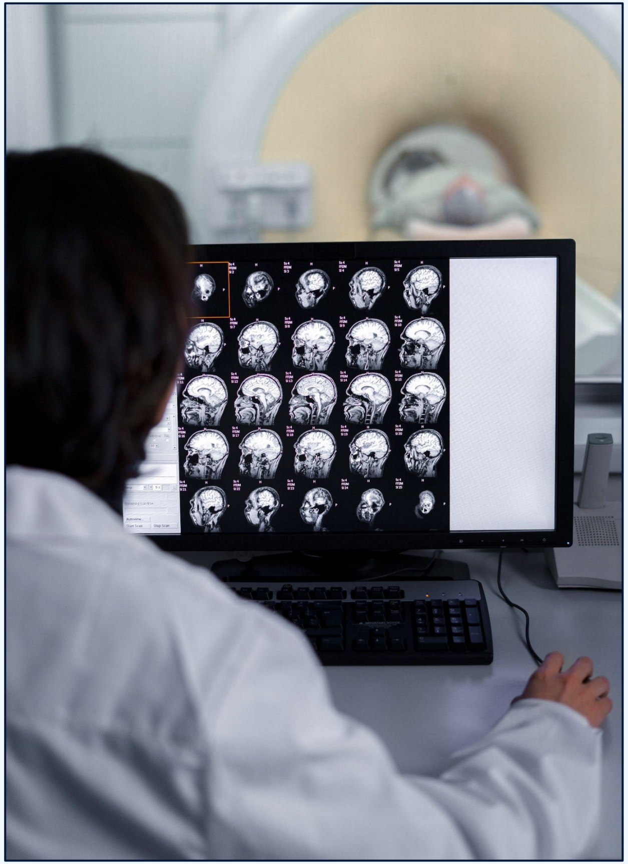
LightGBM





Finding

		Before	After
Random Forest	Accuracy	0.95	0.79
	AUC	0.82	0.66
LightGBM	Accuracy	0.93	0.72
	AUC	0.83	0.81

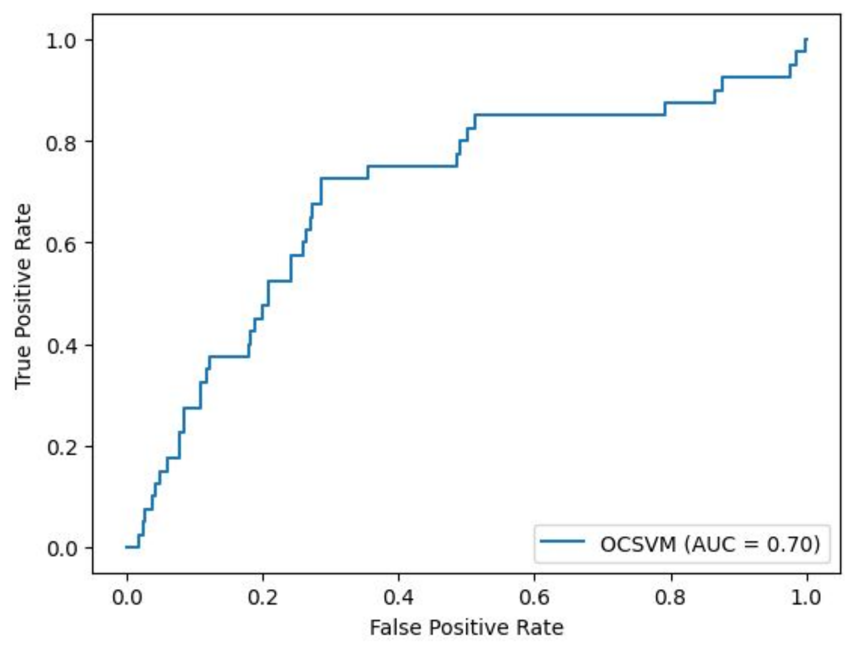


04

Anomaly Detection

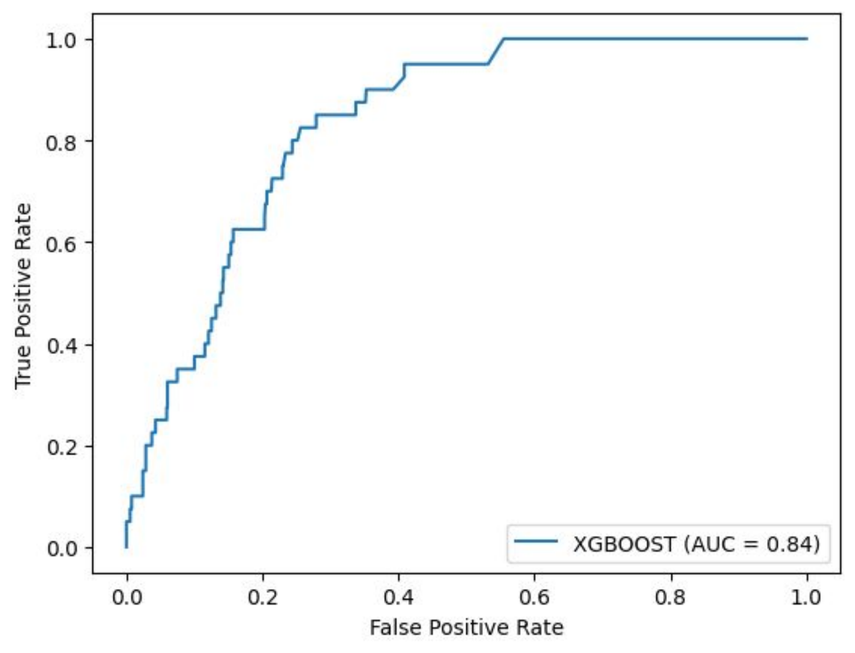
OneClass SVM

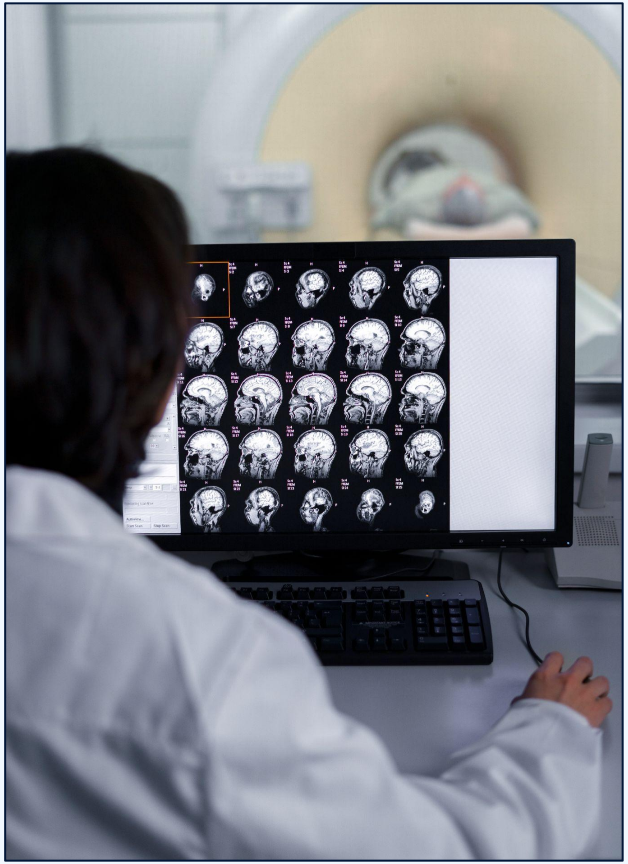
Precision score: 0.1
AUC: 0.7



Extreme Boosting Based Outlier Detection

Precision score: 0.2
AUC: 0.84



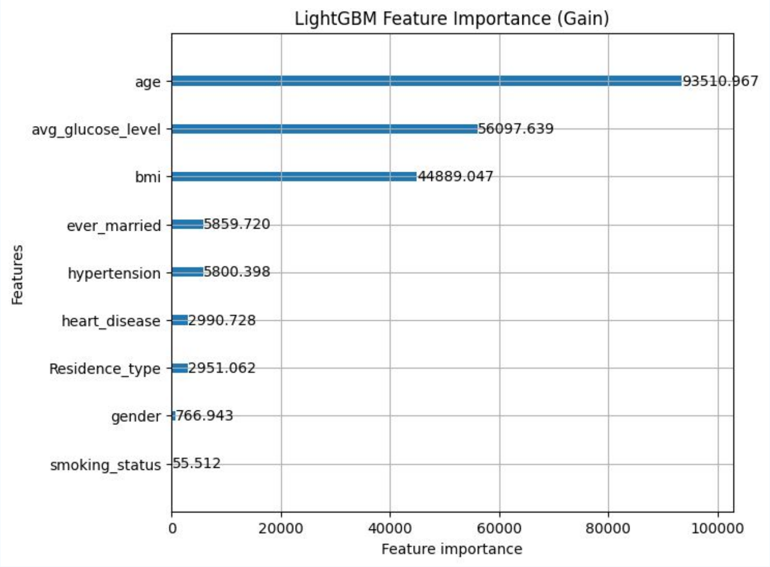


05

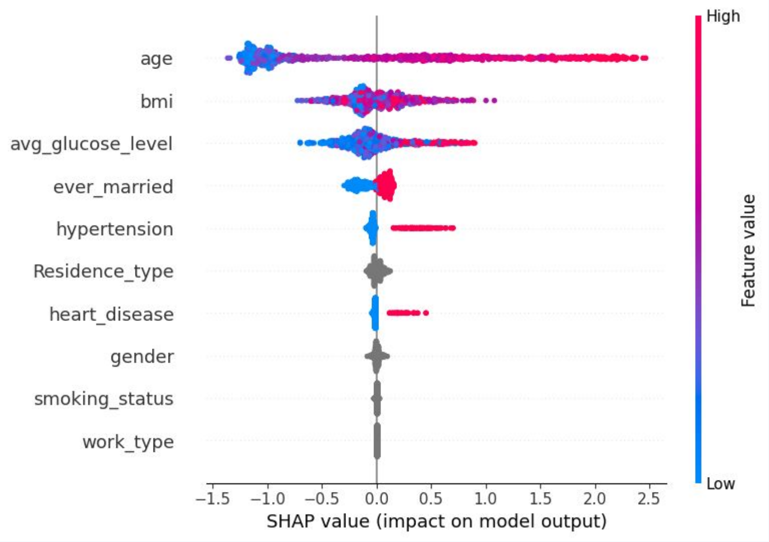
Conclusion

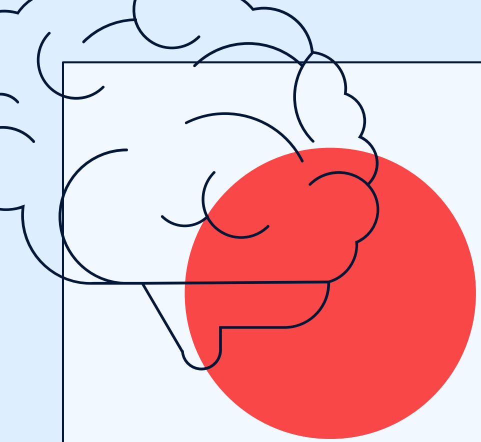
Feature Importance

Gain



SHAP





Thanks

CREDITS: This presentation template was created by **Slidesgo,** and includes icons by **Flaticon** and infographics & images by **Freepik**