NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BS6200 ESSENTIAL MACHINE LEARNING FOR BIOMEDICAL SCIENCE

Final Project

Epileptic-Seizure Classification

NIYUXIN

14/02/2021

CONTENT

1. Introduction

A seizure is an abnormal electrical discharge of a group of brain cells. It can cause different symptoms, depending on the location of the seizure and the spread of electrical activity through the brain.

Fig. 1 Seizure Status

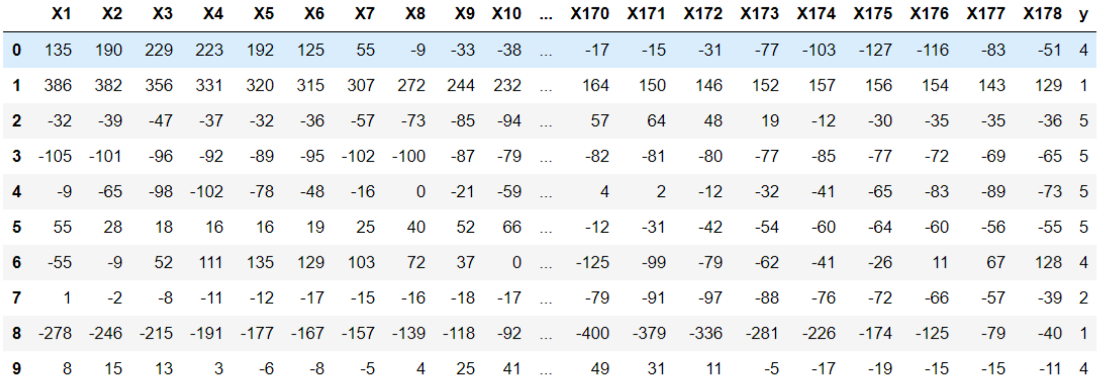
Epilepsy impacts approximately 50 million people worldwide with an estimated annual cost of $12.5 billion for patients in the United States. Epilepsy is characterized by seizures that may impact a person's motor activity with periods of uncontrolled shaking, and are often linked with changes in heart and respiratory rates. In this case, it is important to predict seizures to early recognize.

2. Data Exploration

2.1 Dataset explanation

The original dataset from the reference consists of 5 different folders, each with 100 files, with each file representing a single subject/person. Each file is a recording of brain activity for 23.6 seconds. The corresponding time-series is sampled into 4097 data points. Each data point is the value of the EEG recording at a different point in time. So, we have total 500 individuals with each have 4097 data points for 23.5 seconds.

We divided and shuffled every 4097 data points into 23 chunks, each chunk contains 178 data points for 1 second, and each data point is the value of the EEG recording at a different point in time. So now we have 23 x 500 = 11500 pieces of information(row), each information contains 178 data points for 1 second(column), the last column represents the label y {1,2,3,4,5}, which shows in Fig. 2



	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	y
0	135	190	229	223	192	125	55	-9	-33	-38	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	4
1	386	382	356	331	320	315	307	272	244	232	...	164	150	146	152	157	156	154	143	129	1
2	-32	-39	-47	-37	-32	-36	-57	-73	-85	-94	...	57	64	48	19	-12	-30	-35	-35	-36	5
3	-105	-101	-96	-92	-89	-95	-102	-100	-87	-79	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	5
4	-9	-65	-98	-102	-78	-48	-16	0	-21	-59	...	4	2	-12	-32	-41	-65	-83	-89	-73	5
5	55	28	18	16	16	19	25	40	52	66	...	-12	-31	-42	-54	-60	-64	-60	-56	-55	5
6	-55	-9	52	111	135	129	103	72	37	0	...	-125	-99	-79	-62	-41	-26	11	67	128	4
7	1	-2	-8	-11	-12	-17	-15	-16	-18	-17	...	-79	-91	-97	-88	-76	-72	-66	-57	-39	2
8	-278	-246	-215	-191	-177	-167	-157	-139	-118	-92	...	-400	-379	-336	-281	-226	-174	-125	-79	-40	1
9	8	15	13	3	-6	-8	-5	4	25	41	...	49	31	11	-5	-17	-19	-15	-15	-11	4

Fig. 2 data set

The response variable is y in column 179, the Explanatory variables X1, X2, …, X178

y contains the category of the 178-dimensional input vector. Specifically y in {1, 2, 3, 4, 5}:

5 - eyes open, means when they were recording the EEG signal of the brain the patient had their eyes open

4 - eyes closed, means when they were recording the EEG signal the patient had their eyes closed

3 - Yes they identify where the region of the tumor was in the brain and recording the EEG activity from the healthy brain area

2 - They recorder the EEG from the area where the tumor was located

1 - Recording of seizure activity

All subjects falling in classes 2, 3, 4, and 5 are subjects who did not have epileptic seizure. Only subjects in class 1 have epileptic seizure.

2.2 EEG Distribution

The EEG distribution of different labels are show in Fig. 3, it can be seen that the EEG in seizure is very different to others, the vibration of EEG is extremely severe, then if the violin plot with the 178 mean values is also shown in Fig. 4, the distribution of label 2,3,4,5 is stable, there range is all under 10, however, the mean values of EEG in seizure is approximately 50, and the data is sparce than others, which also shows that the EEG in seizure is very different.

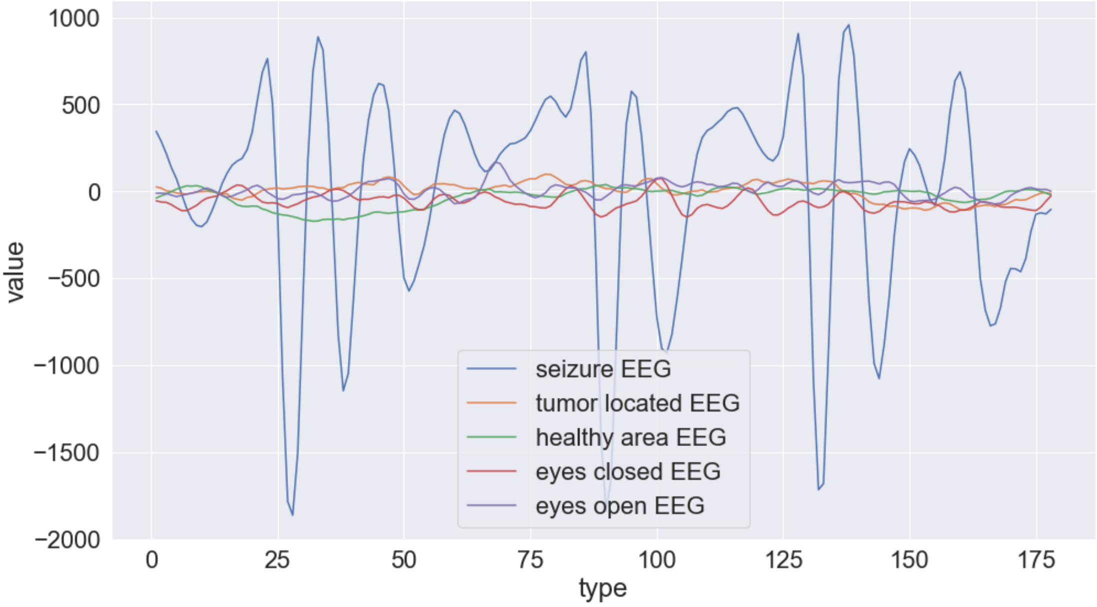


Fig. 3 EEG distribution

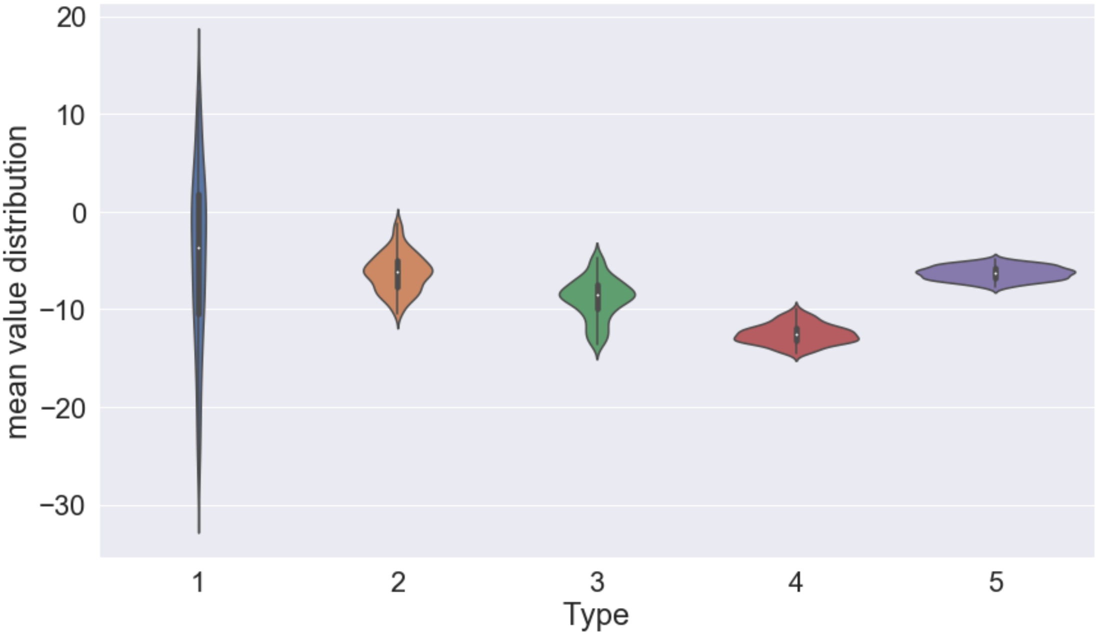


Fig. 4 178 mean value distribution

2.3 Data distribution and PCA visualization

The data distribution is show in Fig. 5, we could see that all five data is totally balanced, and the number of each of them is 2300.Then we transfer them into binary distribution shown in Fig. 6, we could see that the people with no seizure is 4 time larger, which means that our binary data set is totally imbalanced.

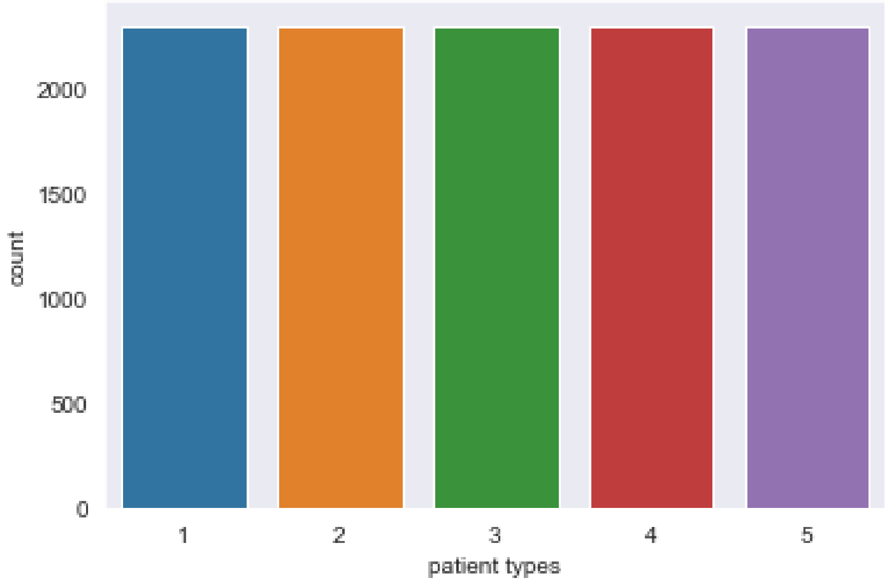


Fig. 5

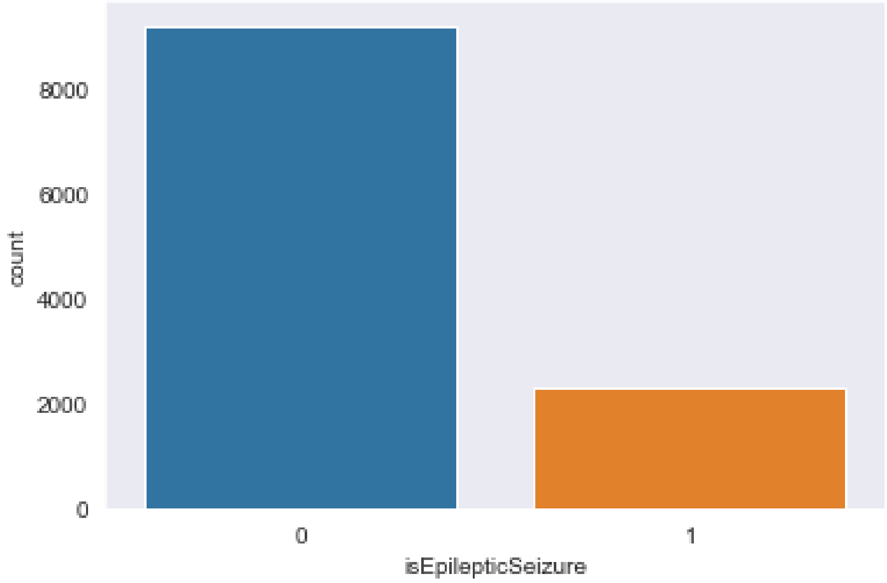


Fig. 6 label distribution

The 2-dimension PCA is also shown Fig. 7, we could see that negative cases are accumulated together and the positive are sparce, but they are overlapping with each other which shows PCA could not cluster then very well and further classification methods should be done.

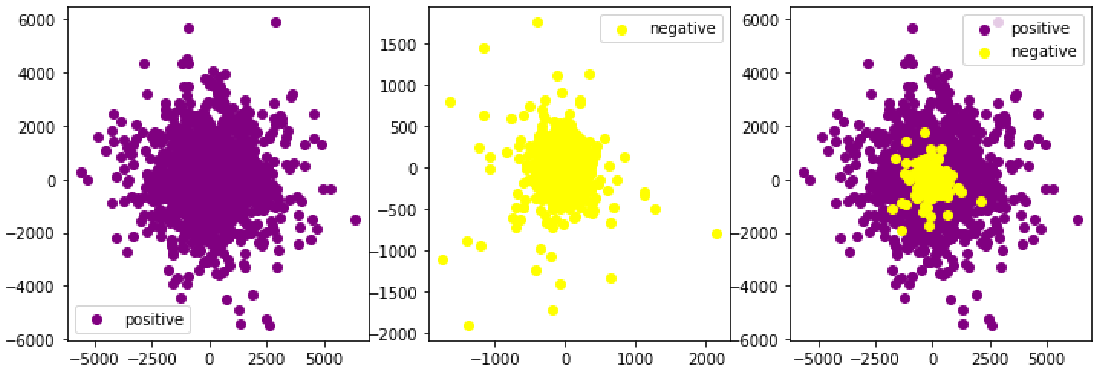


Fig. 7 PCA visualization

3. Preprocessing

3.1 Normalization

From above, we known that the EEG values of seizure and non-seizure is very different, in this cases, normalization should be done to let them in the same range. The standard normalization was used, the formula is shown as below,

$$x_n = \frac{x - \mu}{\sigma}$$

Where μ means the mean of each columns' value, and σ means the standard deviation.

After standard normalization, the distribution of values can be transformed to the distribution with 0 mean and 1 standard deviation and shown in Fig. 8,

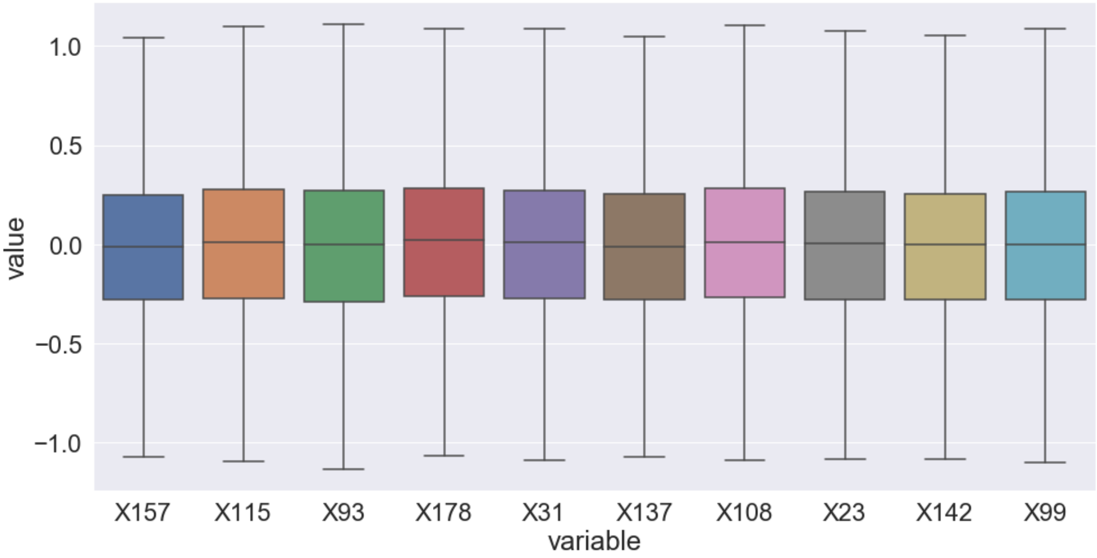


Fig. 8 value distribution after normalization

3.2 Oversampling

Because our binary dataset is highly imbalanced, it is not good for model to train on that kind of dataset, then the data should be oversampling. However, oversampling should not be used in test data, because the test data will be seen by the model if we use the oversampling and let the result cannot be trusted.

In this case, the data set will first be split into 80% train data and 20% train data, then oversampling will be used in train data. Notice that the original data is balanced, so for multi-class classification, there is no need to use oversampling. ADASYN is used to oversample train data, and the train data was increased shown in Fig. 9, we could see that the imbalance is eliminated.

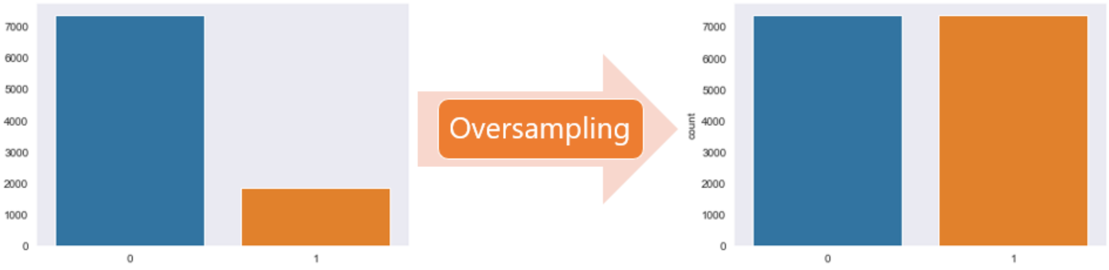


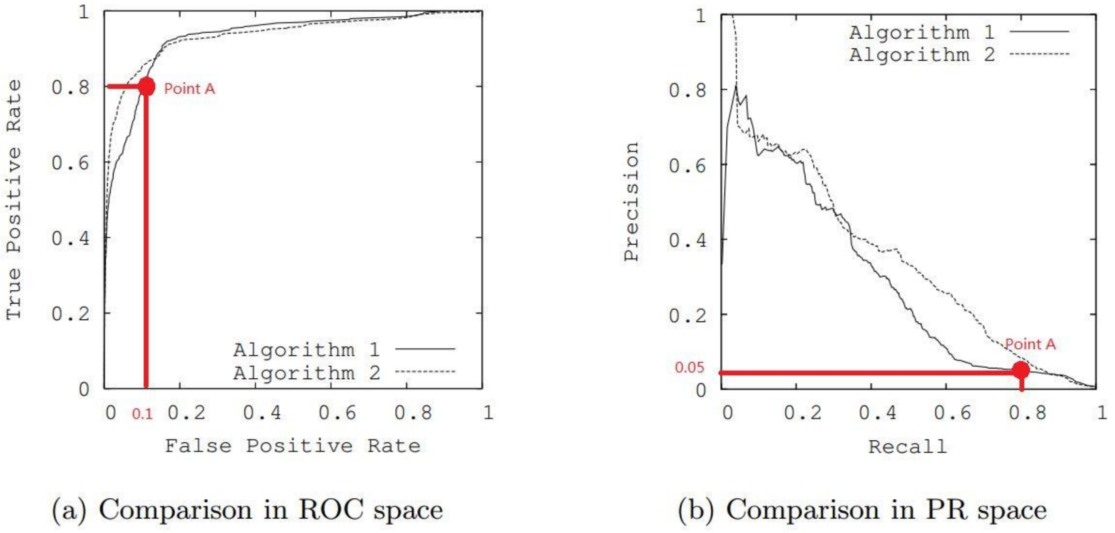
Fig. 9 oversampling

3.3 Metrics selection

3.3.1 Binary Classification metrics

The metrics should be carefully selected towards different predict target. For disease prediction, the number of disease cases is always much less than the number of normal people, which means the dataset is always imbalanced. And our final goal is to predict more accurate towards the patient. In this case, the very popular metrics ROC AUC may be not very suitable for this kind dataset. ROC AUC is unbiased towards model, which means it takes equally to majority class and minority class. This characteristics may let ROC AUC more optimistic when predicting data with highly imbalanced.

David proposed a paper in 2006 in ICML[1] that showed this draw back shown in Fig. 10. The left image is ROC curve, the point more close to left up corner means more better, and the right image is positive data's Presidion-Recall curve, and point more close to right up corner means the point is good. We could see the ROC curve in point A looks very good, the point A is very close to point (0,1), but when we look at the right graph, point A is very far away to the point (1,1), which shows model in this point is poor to predict positive data well.



(a) Comparison in ROC space

(b) Comparison in PR space

Fig. 10 metric choice

In this case, ROC AUC is not suitable, the AUC of Precision-Recall curve is more better for predicting imbalanced data, the PR AUC means the average values of all threshold, so the PR AUC evaluates the model globally in the aspect of positive data, which means it is the suitable tool for us to evaluate a model in the prediction performance toward positive data.

Then in the next page, we will use PR AUC as our first metrics values, and the it is the same to two models, then we will compare there positive F1 score, finally we will compare their positive data recall if the F1 score is also the same.

A threshold to make the positive data F1 score largest is also found by using the PR Curve, the threshold was found by the joint point between the pr curve and the straight line with 1 slope.

3.3.2 Multi-class classification metrics

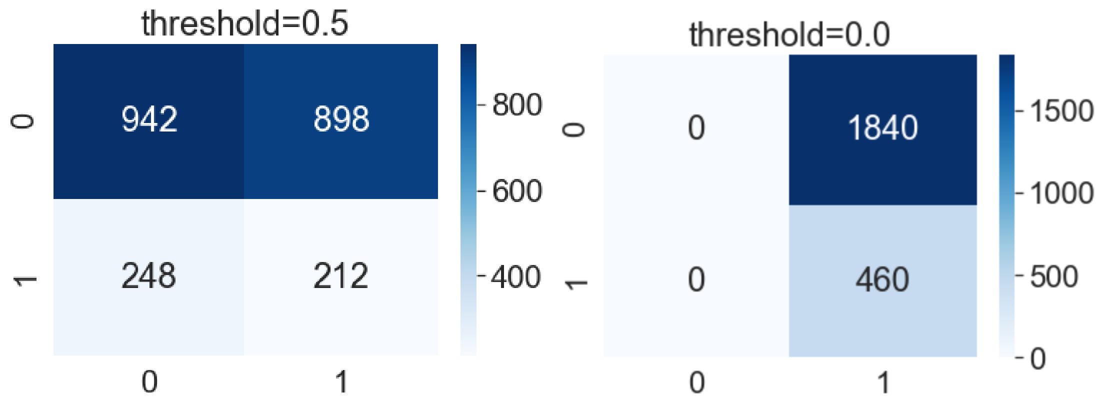
There is no imbalanced data existed in multi-class classification, and the ROC AUC is the global metrics for majority and minority both and AUC is very popular metrics used in the world, then ROC AUC is chosen as the top first metrics in multi-class metrics, then the f1 score will be the second metrics.

4. Model Establishment of Binary Classification

4.1 Base Model

4.1.1 Dummy Classifier

First the dummy classifier is built as our base model, it can be seen that the selection is random chosen. And the best threshold for larges patient (positive data) F1 score is 0, in this case the PRAUC is 0.379.



4.1.2 KNN Classifier

Then KNN is also built in this project, the number of nearest neighbors is chosen based on 5-fold cross validation shows in Fig. 11. The graph shows that PRAUC and ROCAUC increases gradually from 2 to 7 and then start to decrease, the f1 score keeps decreasing monotonously.

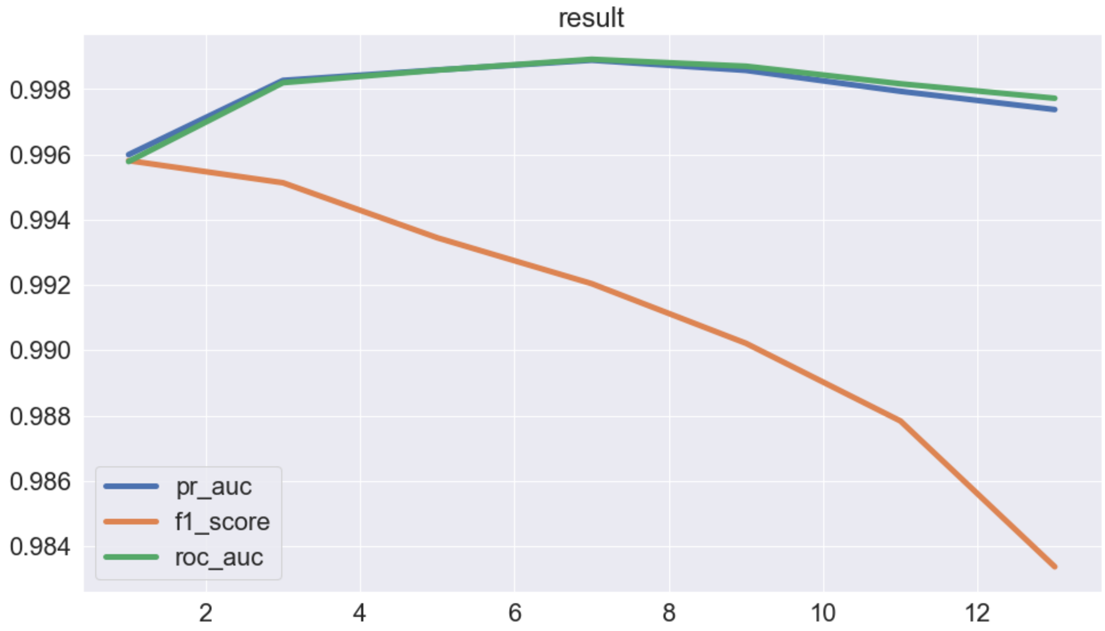
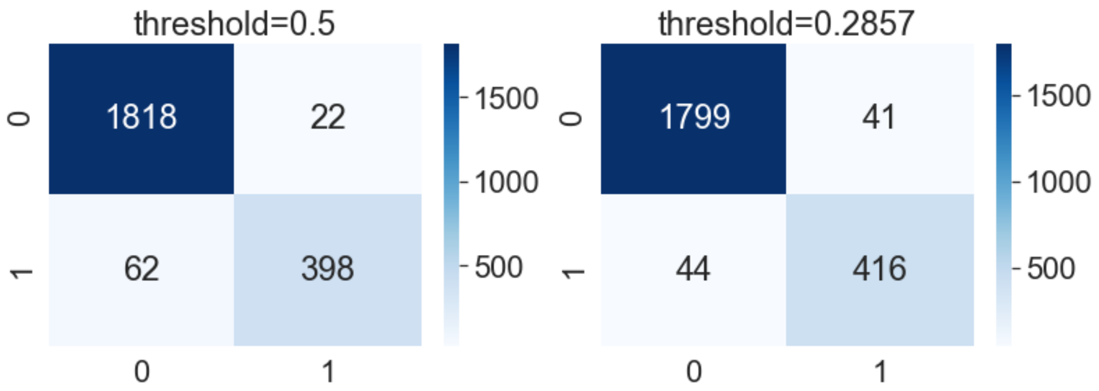


Fig. 11 relationship between number of k and metrics values

Then we choose k = 7 as our final choice after trade-off. The confusion matrix about the test data show below, the PRAUC is 0.95656, best threshold is 0.2857 and the best patient f1 score is 0.907.



4.1.3 Support Vector Machine

SVM is also carried out, there are different kernel for SVM,and it is chosen by 5 fold cross validation again. Fig. 12 shows the different performance of different kernel, it shows that sigmoid has the least performance and rbf kernel has the best, the value of PR AUC and ROC AUC is very close to 1.

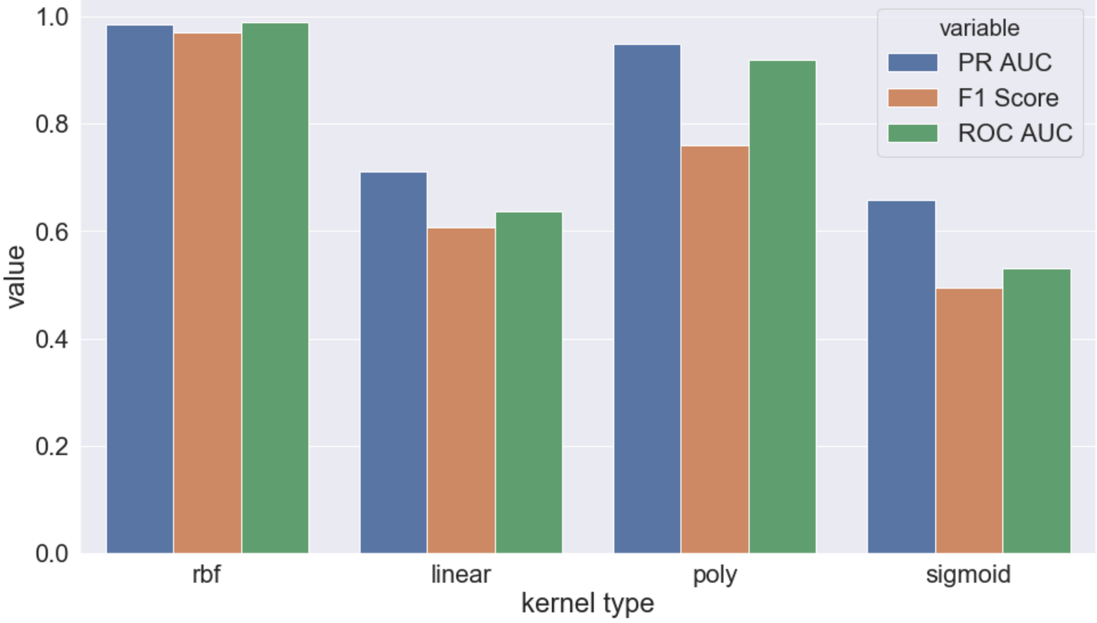
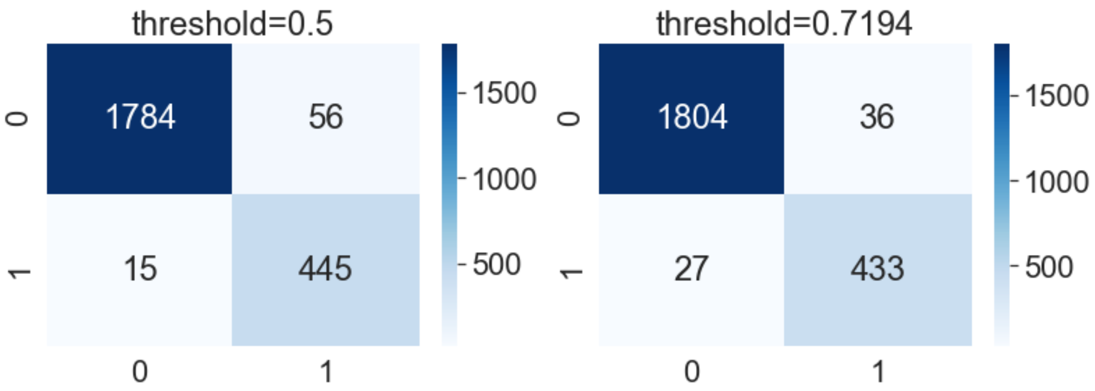


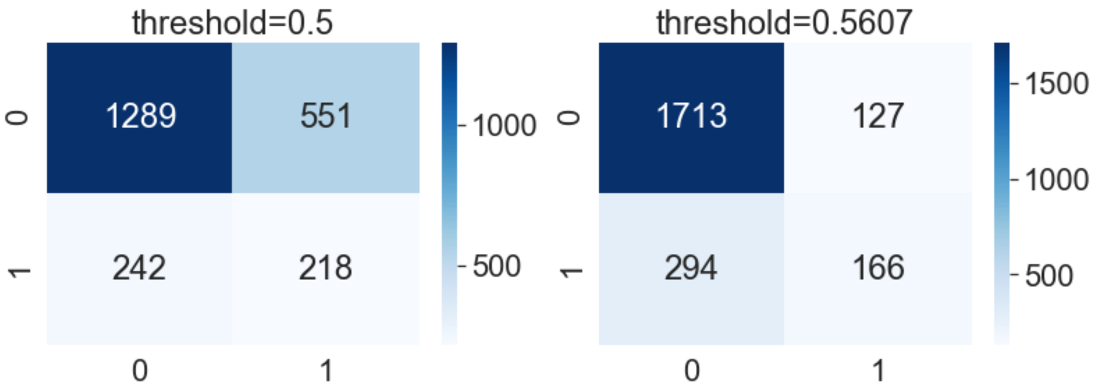
Fig. 12 relationship between type of kernel and metrics values in SVM

Then we choose rbf as our kernel and the performance of this model in test data is show below, the PRAUC is 0.979 which is larger than that of KNN, and the best threshold is 0.7194 with best patient F1 score 0.932.



4.1.4 Logistic Regression

LR is not fine tuned by cross validation, and the performance of this model is shown below, the performance of LR is really bad, the PRAUC is only 0.4378, and the best threshold is 0.5607 with best patient F1 score 0.4409, the f1 score is under 0.5.



4.1.5 Decision Tree

Decision tree model is also built, the hyperparameter is fine tuned by the following metrics values based on 5 fold cross validation at the beginning when the ROC AUC is set as metrics. For saving time, these hyperparameters was used directly. Fig. 13 shows the value change when the max depth changes from 1 to 100, The ROC AUC first at the low level, then jumps to the high value after the max depth larger than 50, and then the value keeps stable, f1 score also increases monotonously. In this case, max depth about 50 was chosen as our final choice.

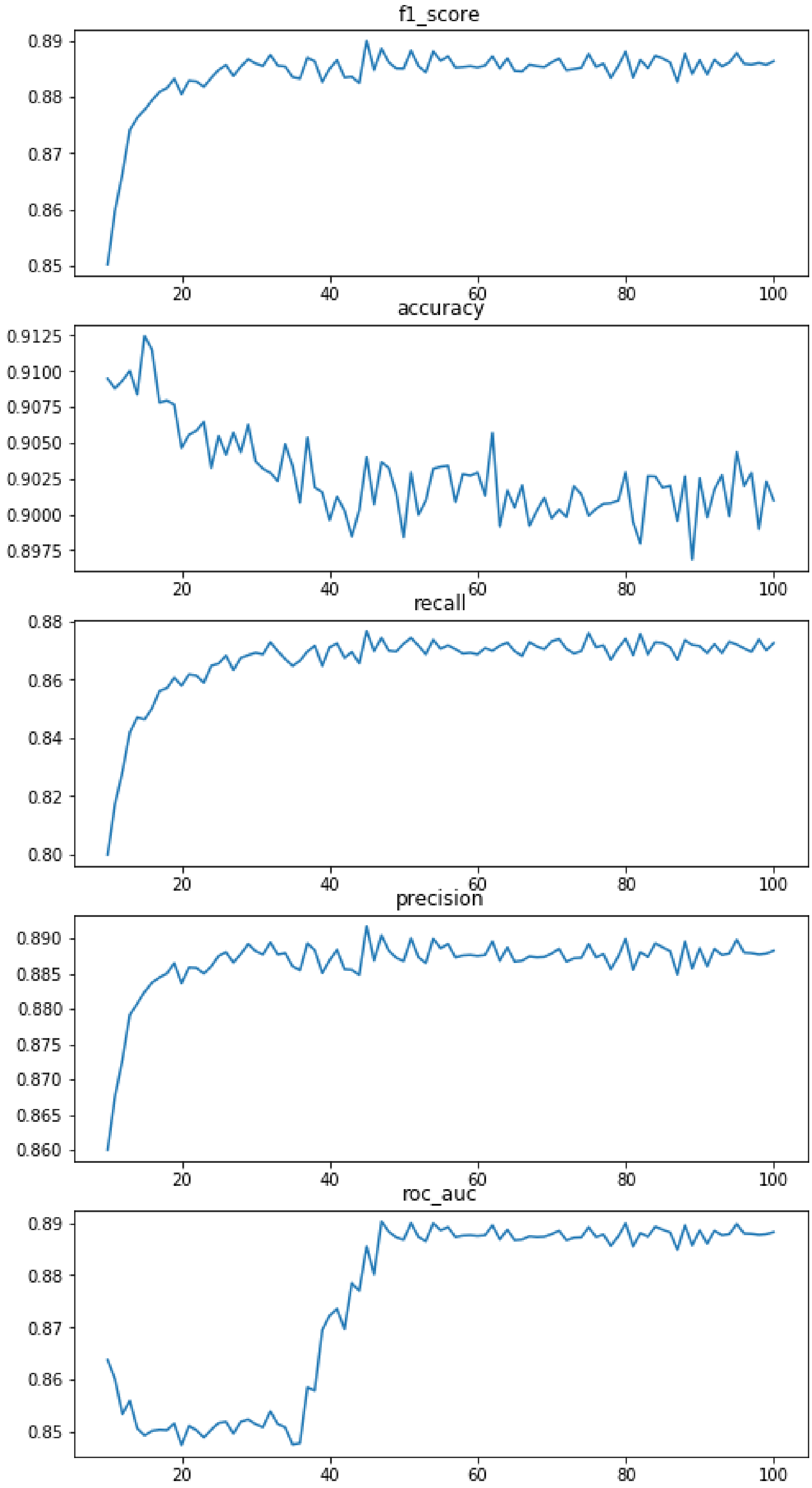


Fig. 13 relationship between max depth and metrics values in Decision tree

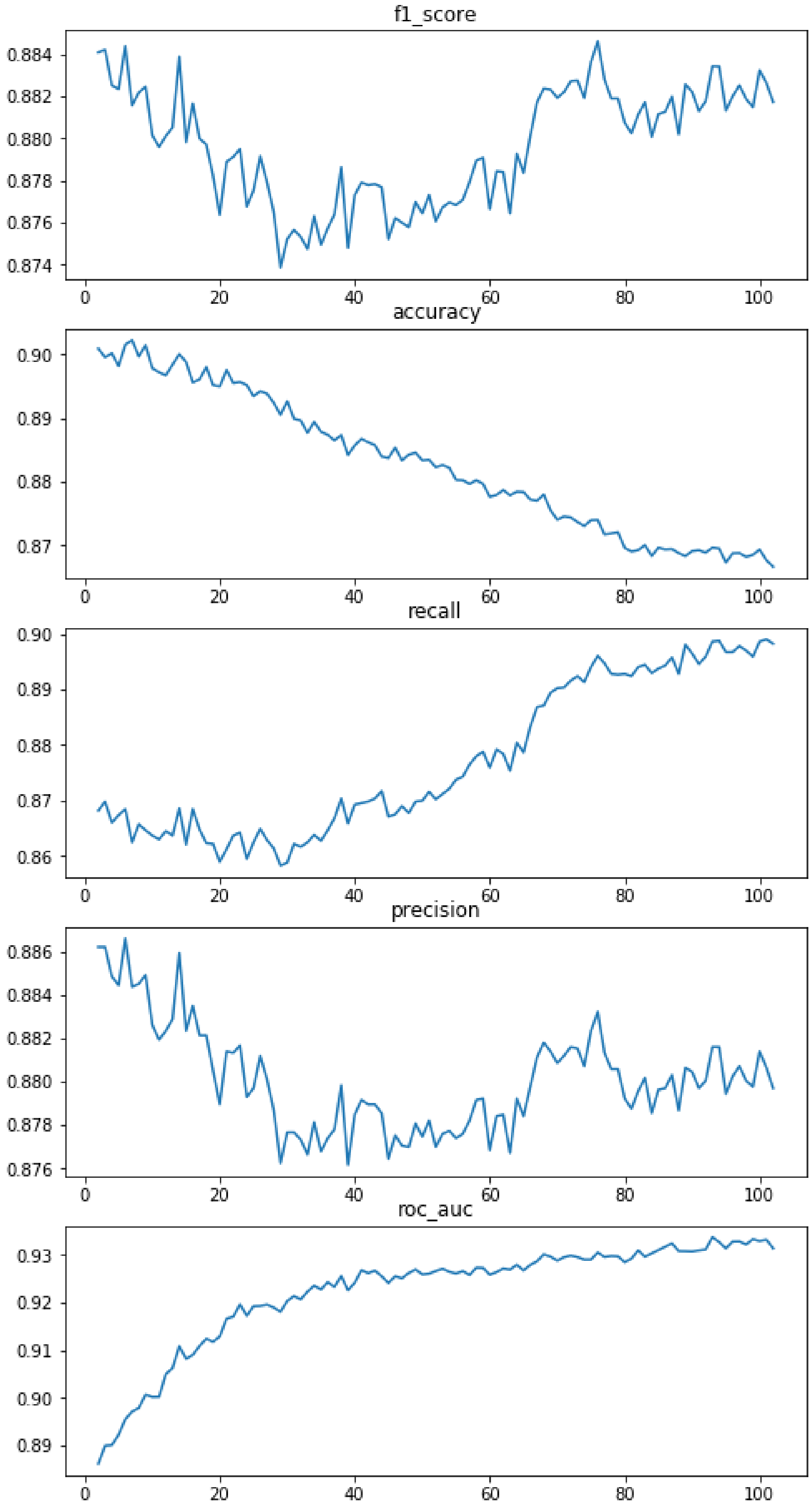
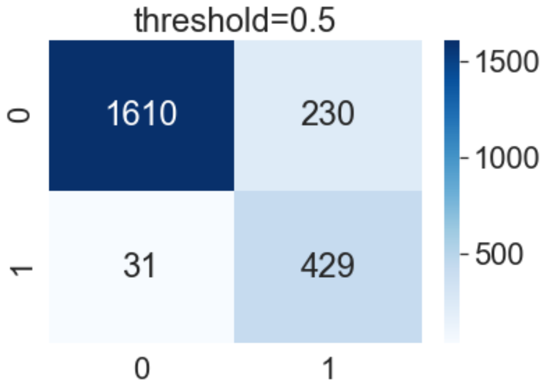
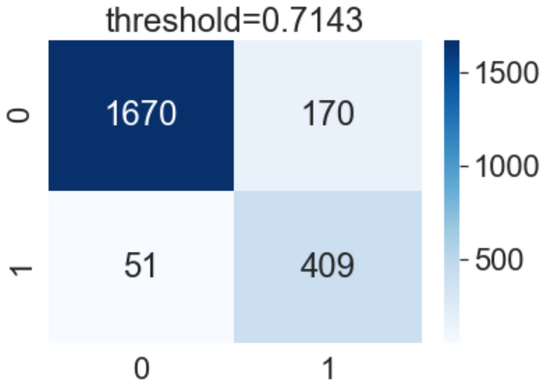


Fig. 14 relationship between metrics value and number of minimum split sample in DT

Fig. 14 also shows the values changes when the minimum split sample change from 1 to 100, the values is not stable like the previous graph. The F1 score keeps decreases from 1 to 30 then starts to increase and ROC AUC increases all the time, then the 100 is chosen after fine tuning.

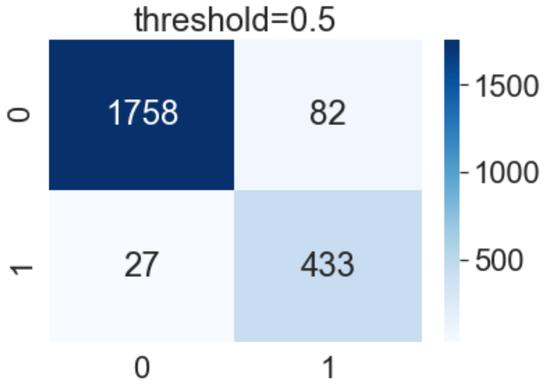
Then the Decision tree with 50 maximum depth and 100 minimum sample split is built as our best decision tree model. The model seems no good at predict normal people, the PRAUC is only 0.849, and the best threshold is 0.7143 with best patient F1 score 0.787, which is also very low but better than Logistic Regression.

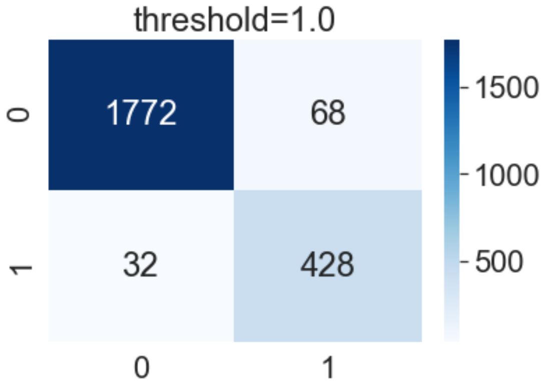




4.1.6 Naïve Bayes

Naïve Bayes is not fine-tuned by the validation part. And the performance of this model is shown below, the performance of Naïve Bayes is good, the PRAUC is 0.9227,and the best threshold is 1.0 with best patient F1 score 0.8954.



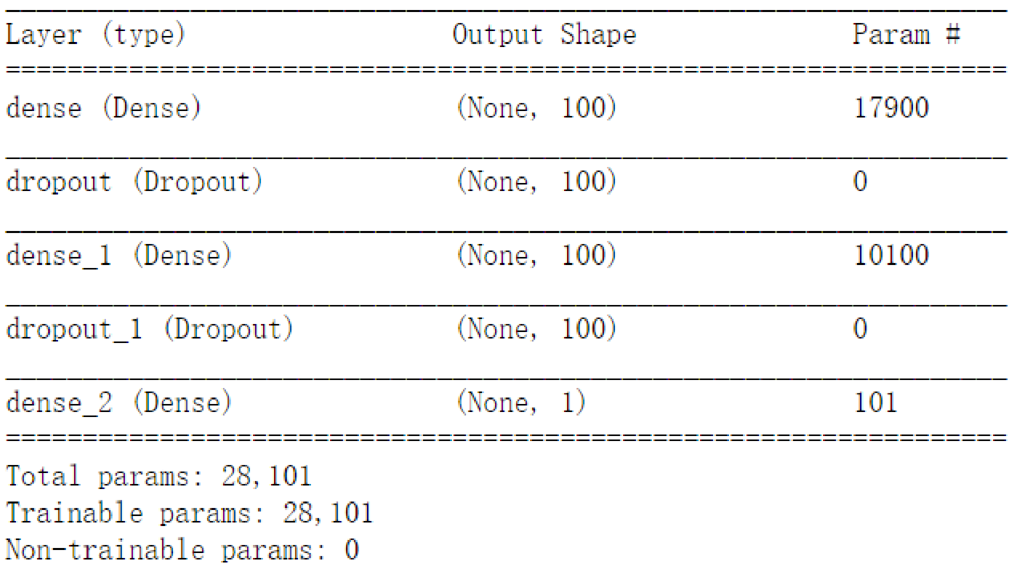


4.1.7 Multi-Layer Perceptron

There two different ways to implement MLP method, the first is to use deep learning method and the second thing is to use the MLP method encapsulated by sklearn. All these two methods are carried out to build a MLP method.

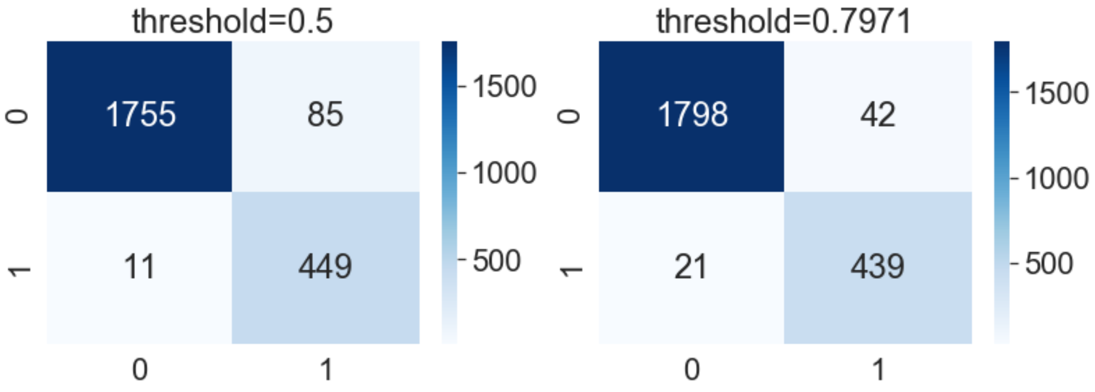
4.1.7.1 4-Layer DNN

The 4 layer DNN contains 1 input layer, two hidden layer and 1 output layer, the model structure is shown in Fig. 15. The number of neural of hidden layers are both 100 with 0.5 dropout, and then a sigmoid function is used as final output value.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 100)               17900
_________________________________________________________________
dropout (Dropout)            (None, 100)               0
_________________________________________________________________
dense_1 (Dense)              (None, 100)               10100
_________________________________________________________________
dropout_1 (Dropout)          (None, 100)               0
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 101
=================================================================
Total params: 28,101
Trainable params: 28,101
Non-trainable params: 0
```

Fig. 15 4-Layer structure



The confusion of 4-layer DNN are shown below, the PRAUC is 0.9803, and best patient F1 Score is 0.933 with best threshold 0.7971.

4.1.7.2 MLP

MLP is carried out by the sklearn tools, the number of hidden layers are fine-tuned by 5 fold cross validation. And the best performance of these model is 0.9433 with number 256 of neural perceptron in hidden layer.

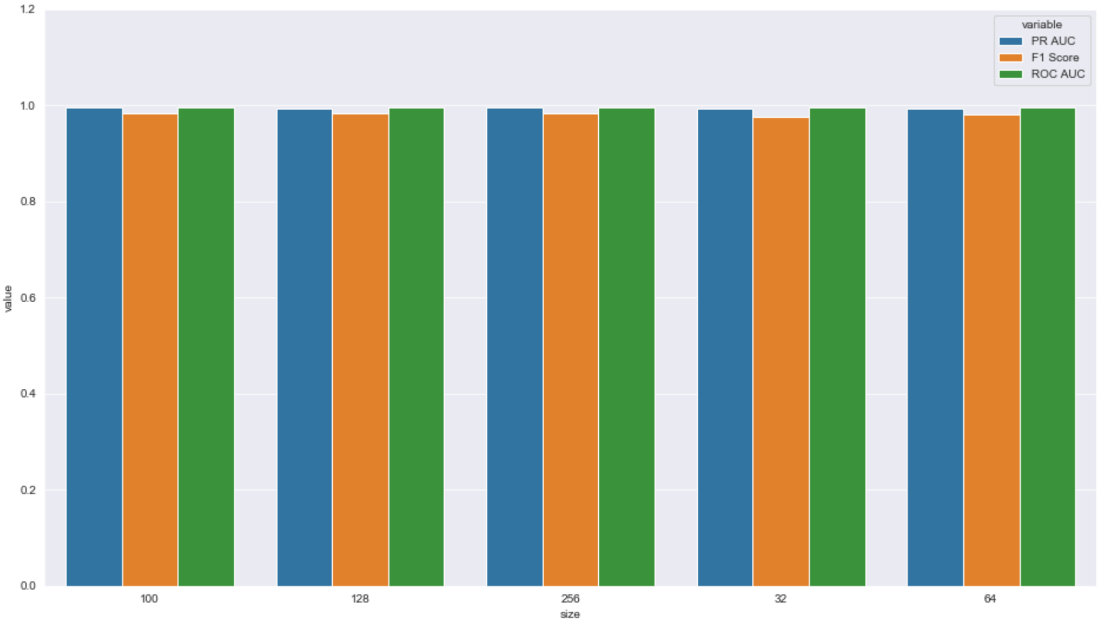
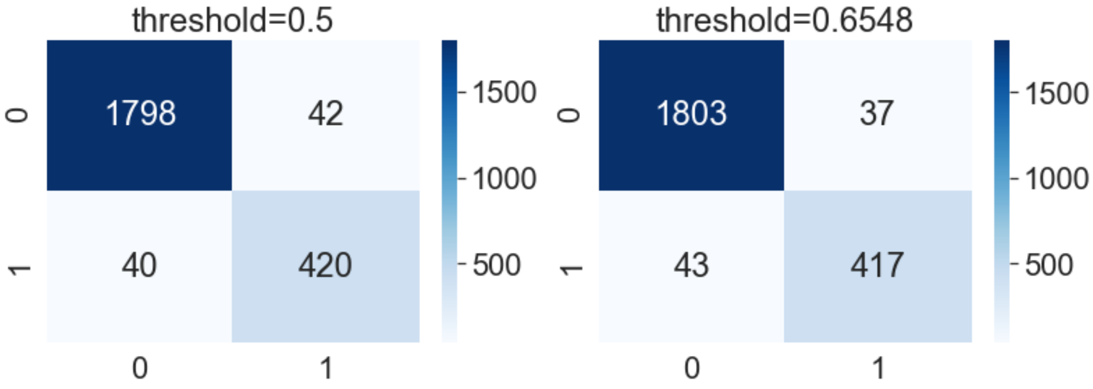


Fig. 16 relationship between type of number of hidden layer and metrics values in MLP

In this case, the final MLP with two 256 hidden layers are used for prediction and confusion matrix are shown below, the MLP model seems also predict well. The PR AUC is 0.9598, and the best patient F1 score is 0.912 with threshold 0.6548.



4.2 Ensemble Learning

4.2.1 Random Forest

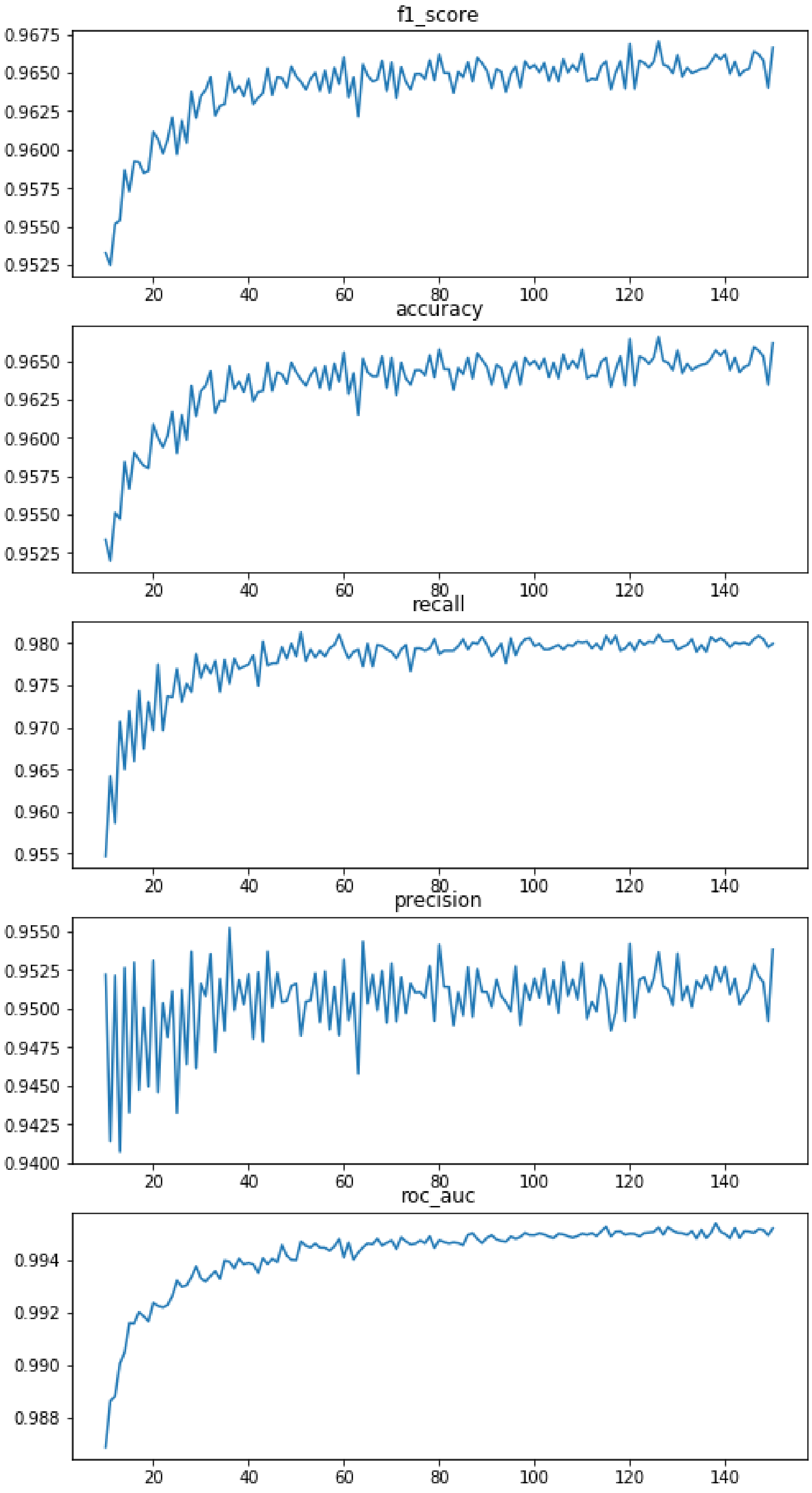


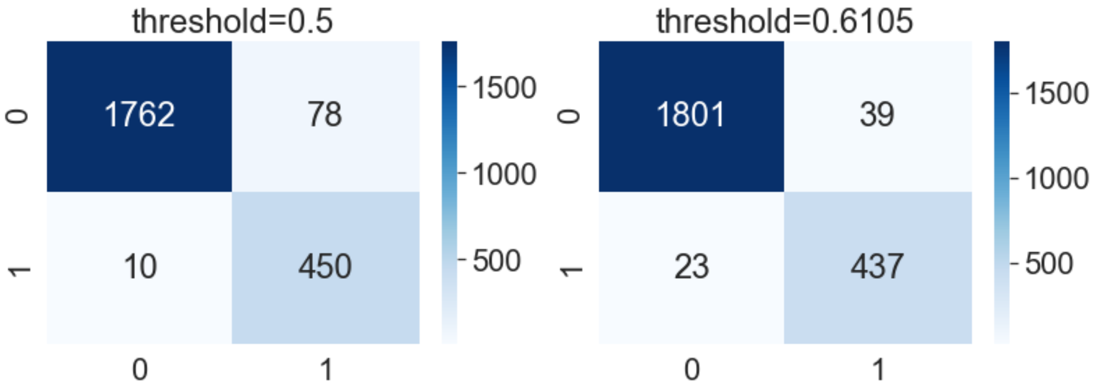
Fig. 17 relationship between number of n estimators and the metric values

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

The hyperparameter (n estimators and min samples split) of random forest is fine tuned by the 5-fold cross validation.

Fig. 17 shows the model performance of n-estimators with different values, the model shows that when the number increases, the performance model increases monotonously but the performance does not change much after 100, In this case, 100 is chosen as n estimators. And 4 of min sample split is also selected from the range of 2,4,6,8,16,32.

In this case the fine tuned Random forest is established and the confusion matrix is shown below, the PRAUC is 0.97915, and best patient F1 score is 0.9295 with best threshold 0.6105.



4.2.2 XGBoost

XGBoost is also an ensemble learning method, it is a widely used supervised learning algorithm based on GBDT, it was trained by boosting tree and decision tree. The advantage of XGBoost is effective, efficient, and suitable for parallel calculation.

The n estimators, max depths and gamma coefficient is fine tuned by the 5 fold cross validation. Fig. 18 shows the relationship between the performance and the number of n estimators. Like the random forest, performance also increases when the number of n estimator increases. 10 is chosen in range of 4,6,8,10,12,24, as max depth and 0 is chosen as the gamma in the range of 0, 0.1, 0.2, 0.3, 0.4, 0.5,0.6 since the PRAUC is the best between them.

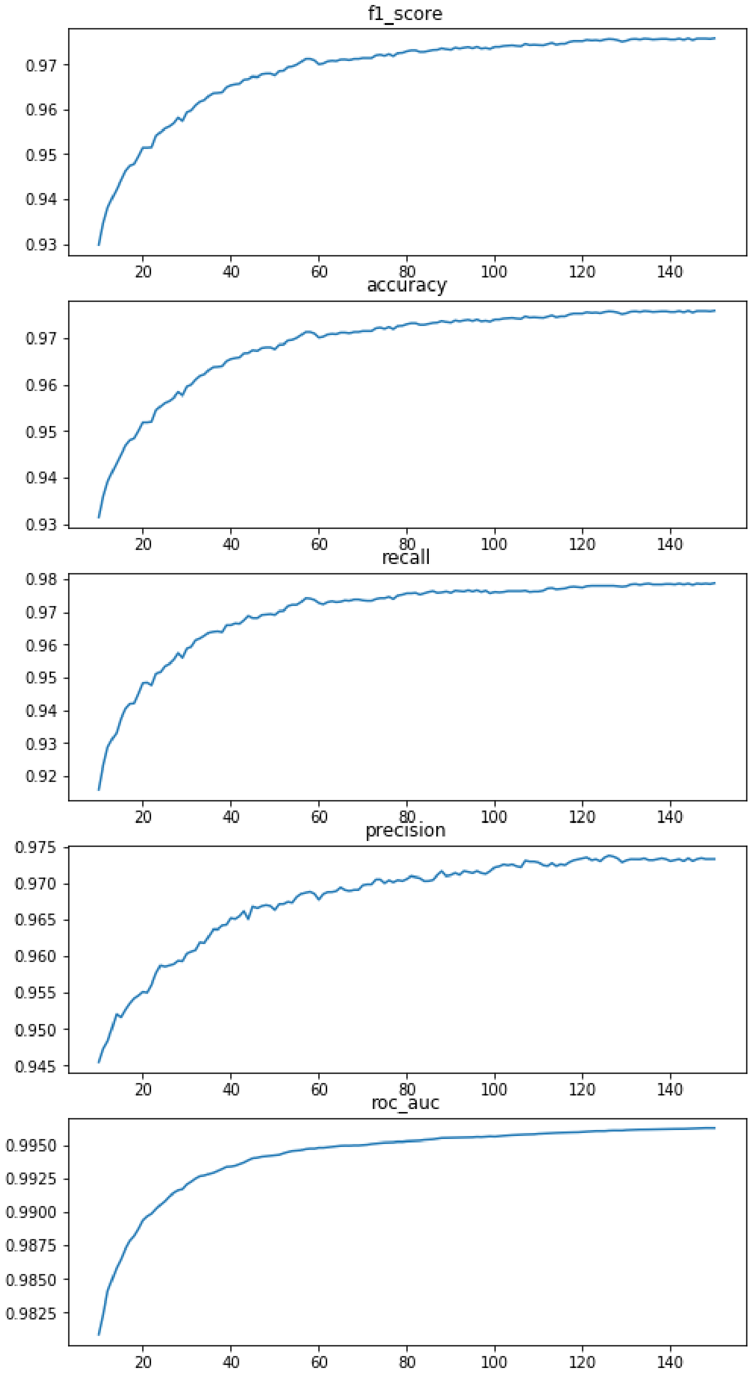
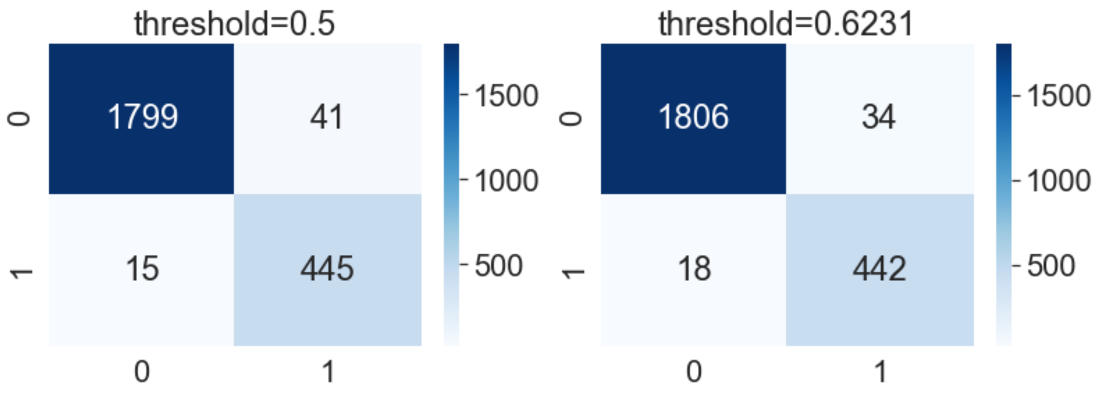


Fig. 18 relationship between n-estimators and metric values in XGBoost

After the fine tune, the XGBoost with n-estimator =150, max depth=10 and gamma =0 is selected and the confusion matrix is shown below, the PRAUC is 0.98560 and best patient F1 score is 0.94444 with best threshold 0.6231.



4.2.3 Voting Classifier

Voting Classifier is the tool in sklearn to combine conceptually different machine learning classifiers and use a majority vote or the average predicted probabilities (soft vote) to predict the class labels. Herein, the soft voting is selected.

Since every base model has different performance, a weight will used to let the voting classifier pay much attention on the prediction on the best model. The formula of attention weight is as shown below,

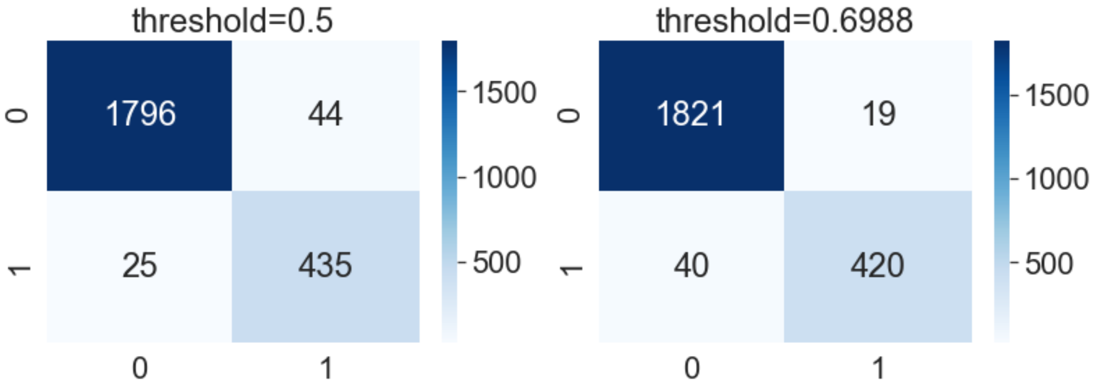
$$Weight = \frac{Test\_data\_PR\_AUC}{1 - Test\_data\_PR\_AUC}$$

Where test_data_PR_AUC is the PRAUC towards the test data, this formula will make the bigger values much bigger and help the model focus on the model with the best performance in test data.

4.2.3.1 Voting with base model

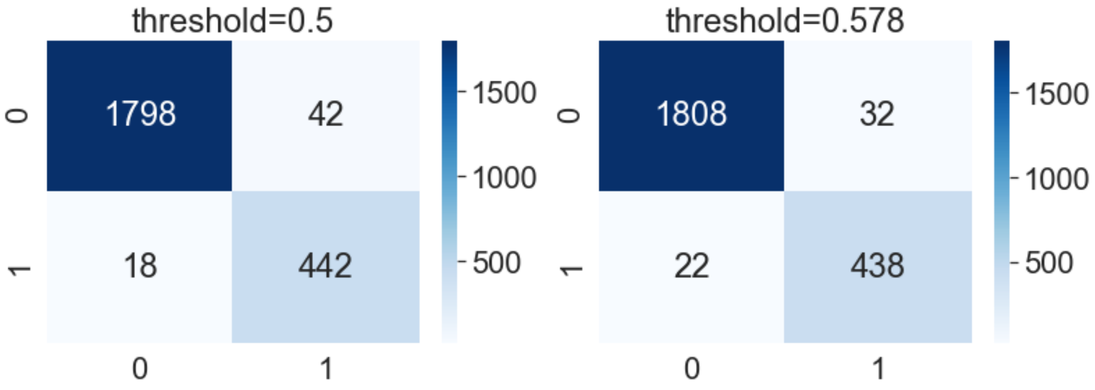
The voting with base model is established, however, Dummy Classfier, Logistic Regression are eliminated because their performance are not good. 4-layer-DNN is also eliminated because model established by tensor-flow is not compatible with voting classifer.

The other model are all selected. and the weight is 4， 8， 1， 2， 4 to the model KNN,SVM, Decision Tree, Naïve Bayes, MLP, respectively. The final confusion matrix is shown below, PRAUC is 0.98448, and best F1 score is 0.9344 with best threshold 0.6988.



4.2.3.2 Voting with all models

Finally, all the previous models including random forest and XGBoost are established, and the weight is 4,8,1,2,4,9,11 to the model KNN, SVM, Decision Tree, Naïve Bayes, MLP, Random forest and XGBoost respectively. The confusion matrix is shown below, PRAUC is 0.98878 and best patient f1 score is 0.9419 with best threshold 0.578.



4.3 Model summary

The following Table 1 shows the performance of all the models except Dummy and 4-layer-DNN. We could see that in the base models, PR AUC and ROC AUC of SVM is largest. However, all ensemble learning method are better than base models. Voting with base models are better than random forest but not better than XGBoost. And voting with all models has the largest PRAUC and ROCAUC. It also can be found that all the average F1 score of the best threshold is larger than average weighted F1 score with 0.5 as threshold. And the patient F1 score is still the largest one, better than voting with all models.

Table 1

Model	Fine tune?	hyperparameter	Best threshold for patient	Average Weighted F1 Score		Patient F1 (best)	ROC AUC	PRAUC
				threshold=0.5	threshold =best threshold			
KNN	Y	k=7	0.2857	0.9628	0.963	0.907	0.958	0.957
SVM	Y	kernel="rbf”	0.7194	0.970	0.9722	0.932	0.994	0.979
LR	N		0.5607	0.6828	0.8	0.441	0.528	0.438
Decision Tree	Y	max depth=50, min sample split=100	0.7143	0.8919	0.9069	0.787	0.9485	0.849
Naïve Bayes	N		1	0.9536	0.9566	0.8954	0.9786	0.92271
MLP	Y	hidden layers=(256,256)	0.6548	0.9644	0.9651	0.9124	0.9774	0.95982
Random Forest	Y	n_estimators=100, min_samples_split=4	0.5717	0.9619	0.9712	0.930	0.9953	0.9792
XGBoost	Y	n_estimators=150, max_depth=4,gamma=0	0.6231	0.9759	0.9771	0.9444	0.9962	0.9856
Voting-with-base-model-1	N		0.7865	0.9699	0.9728	0.9343	0.9957	0.9845

Voting-with-all-models-2	**N**		**0.578**	**0.97416**	**0.9766**	**0.9419**	**0.9970**	**0.98878**

The PR Curve is also shown in Fig. 19, it is better if the curve close to the right up corner and the joint point with the slope 1 straight line is the best F1 score point. We could see that logistic regression is low, other models are all good and all close to right up corner.

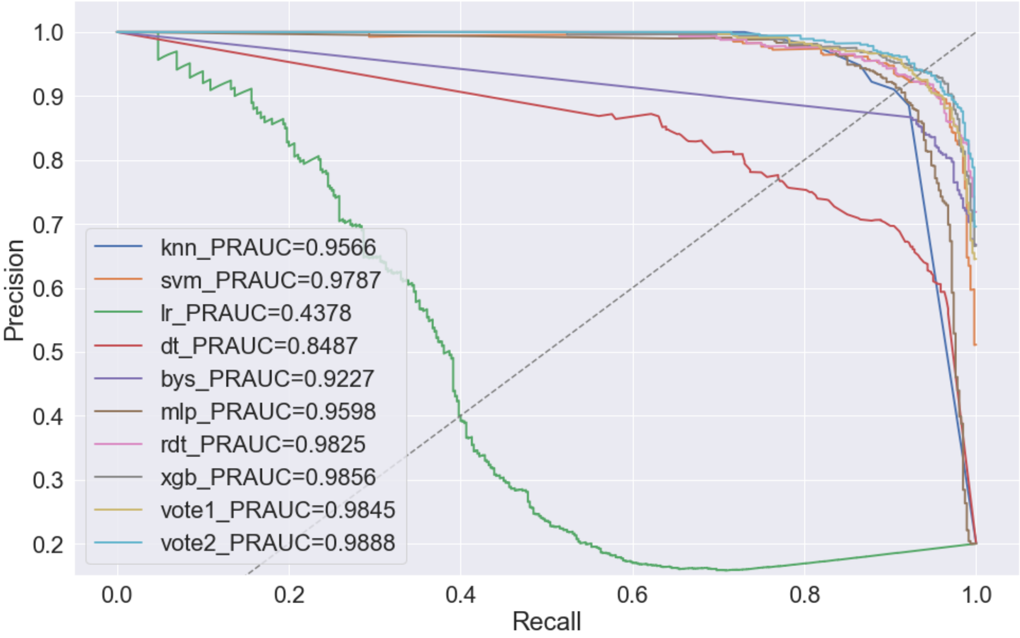


Fig. 19 PR AUC ditribution

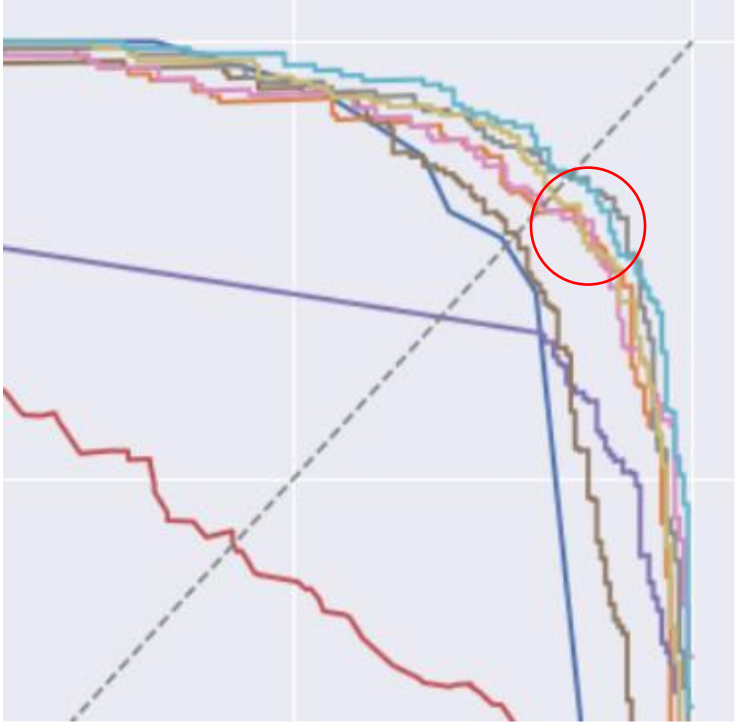


Fig. 20 zoomed graph in right up corner

For more detail, the zoomed graph is shown in Fig. 20 shows the detail about the reason why f1 score of XGBoost is larger than voting classifier with all models. The Blue line is voting classifier with all model and gray line is XGboost, we could see in the red circle. XGboost is more close right up corner, and the joint point is in that circle. However, the performance of voting classifier which is outside of the circle is much larger better than that of in XGBoost.

Because PRAUC is the global metrics for minority class, the voting classifier with all models are chosen as the final model for binary classification.

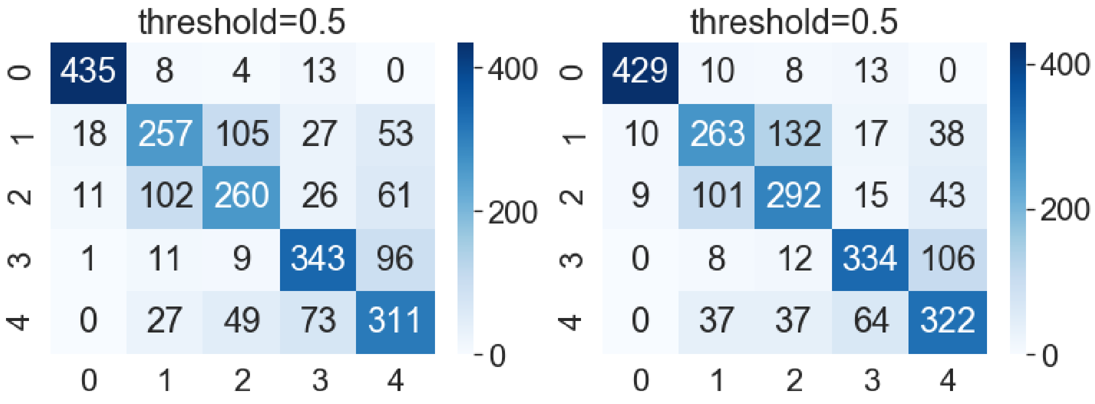
5. Model Establishment of Multi-Class Classification

Multi-class classification is also carried out by using ensemble learning models. Since there is no imbalanced data exist in multi-class, we use the average ROC as our metrics. The Table 2 shows that the performance of voting with all models still is the best model for multi-class classification and the XGBoost is also the second based on ROCAUC. And it can be seen that it is much difficult for machine to predict multi-class model, the ROCAUC and F1 score is much lower than that of in binary classification.

Table 2

Model	Average F1 Score	ROCAUC
Random Forest	0.697	0.916
XGBoost	0.714	0.924
Voting-with-base-models	0.724	0.909
Voting-with-all-models	0.736	0.9338

The confusion matrix below shows the detail of the classification when threshold is 0.5, 0,1,2,3,4 means the label 1,2,3,4,5, respectively. It can be seen that type 2 and 3 is very hard to recognized, nearly all models are confused with these two types. Class 4 and class 5 is also very similar, which makes sense because one is the EEG of eyes open and the other is about EEG when eyes close. Patient with seizure can be recognized easily compared to other 4 types.



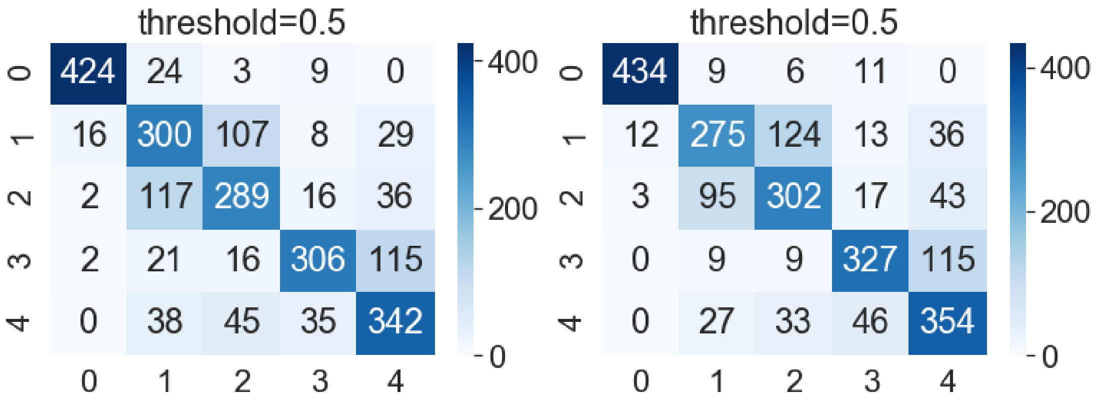
threshold=0.5

	0	1	2	3	4
0	435	8	4	13	0
1	18	257	105	27	53
2	11	102	260	26	61
3	1	11	9	343	96
4	0	27	49	73	311

threshold=0.5

	0	1	2	3	4
0	429	10	8	13	0
1	10	263	132	17	38
2	9	101	292	15	43
3	0	8	12	334	106
4	0	37	37	64	322

Fig. 21 (left)Random Forest,   (right)XGBoost



threshold=0.5

	0	1	2	3	4
0	424	24	3	9	0
1	16	300	107	8	29
2	2	117	289	16	36
3	2	21	16	306	115
4	0	38	45	35	342

threshold=0.5

	0	1	2	3	4
0	434	9	6	11	0
1	12	275	124	13	36
2	3	95	302	17	43
3	0	9	9	327	115
4	0	27	33	46	354

Fig. 22 (left)Voting classifier with base models,   (right)Voting classifier with all models

6. Analysis

6.1 Feature importance in binary classification

To better understanding the result and interpret the model, feature importance is also shown in Fig. 23. The red and blue legend means value of the features, red means the feature is larger and blue is less. Because the best model-voting with all models are mixed models and there is no feature importance, then the second model——XGBoost is chosen to show feature importance.

Fig. 23 shows that X1, X2, and X178 is top 3 features for model to predict whether a patient has epileptic seizures. And we also could see that high value of X1, X2, X178, X96, X131, X159,X86, X32 have more large percentage to have seizure, while low value of feature X164,X126, X51,X39, X50 will cause seizure.

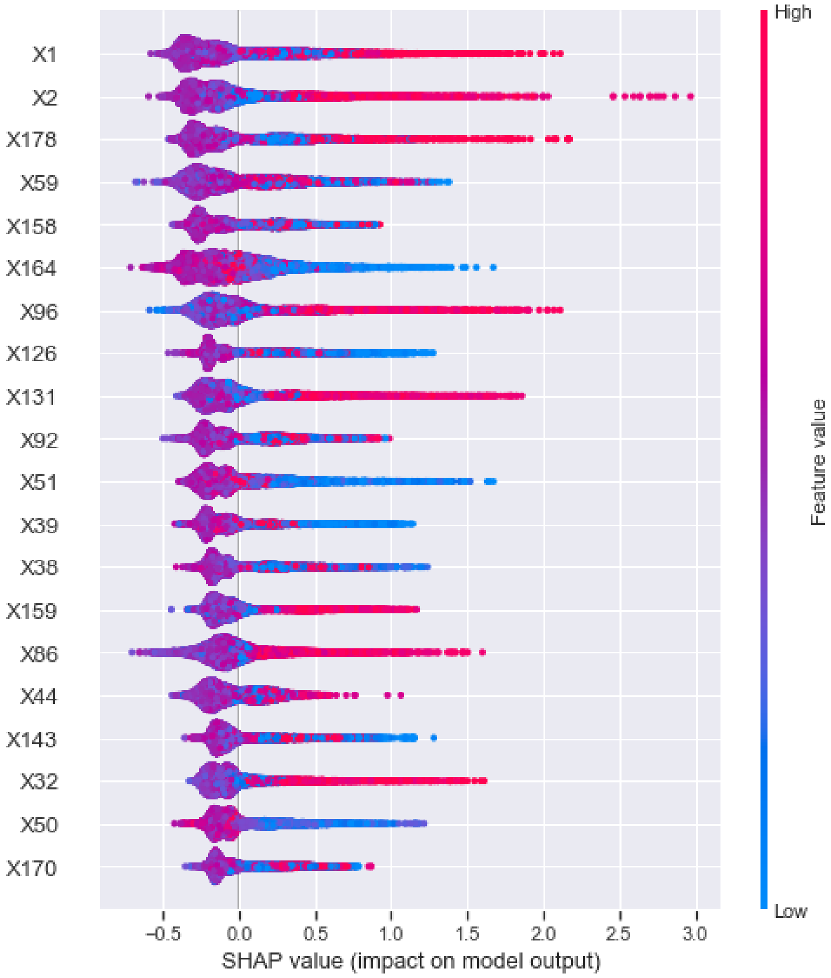


Fig. 23 Feature importance

To be more specific, the relationship between shap value and value of features are selected in Fig. 24. It can be seen that when the value of X2 is larger than 0, the sample will more likely to get seizure, especially in the range from standard value 1 to 2.5 (true range 155.14~404.24), the percentage to get seizure is extremely high. The tendency of X164 is just the opposite and we could see the value of X164 is less than 0, the sample is more likely to get seizure, especially in the range between -2.5 to -1(true range, -427.28~-176.78).

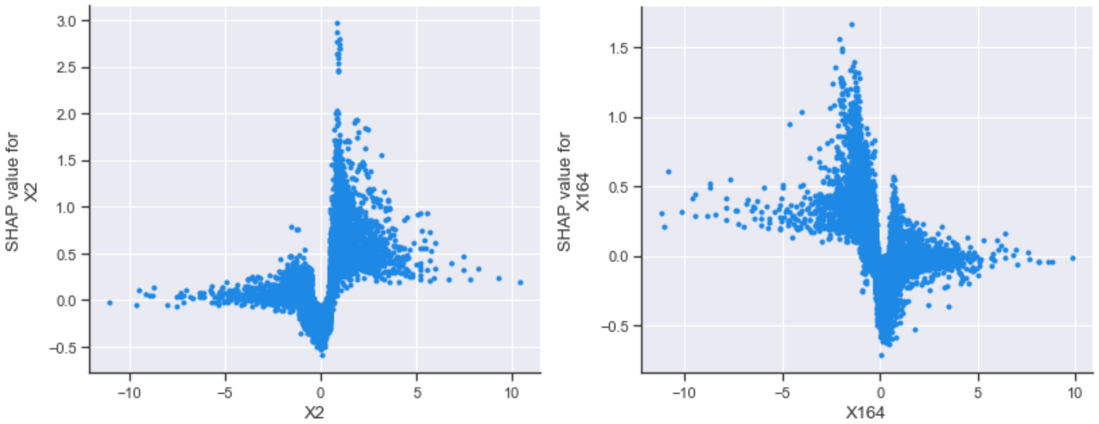


Fig. 24 relationship between specific value and feature importance value in binary classfication

6.2 Feature importance in Multi-Class classification

Fig. 25 shows the feature importance of multi-class classification, X1,X12 and X173 are considered the top 3 important features.

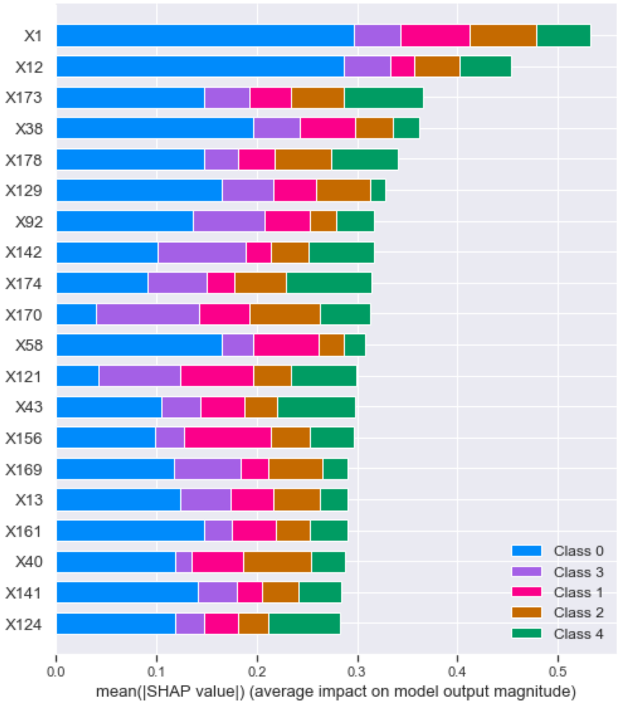


Fig. 25 feature importance of multi-class classification

The following 4 graphs are the feature importance for the 2,3,4,5 labels. The graph shows that X33,X22 and X156 are the top three for label2 (left up corner graph), X49, X170 and X171 are the top 3 for label 3(right up graph), X164, X170 and X98 are the top 3 for label 4 (left bottom graph)and X133, X136 and X114 are the top 3 for label 5 (right bottom graph). The relationship between percentage of the label predicted by model and the value of features are also can be figured out.

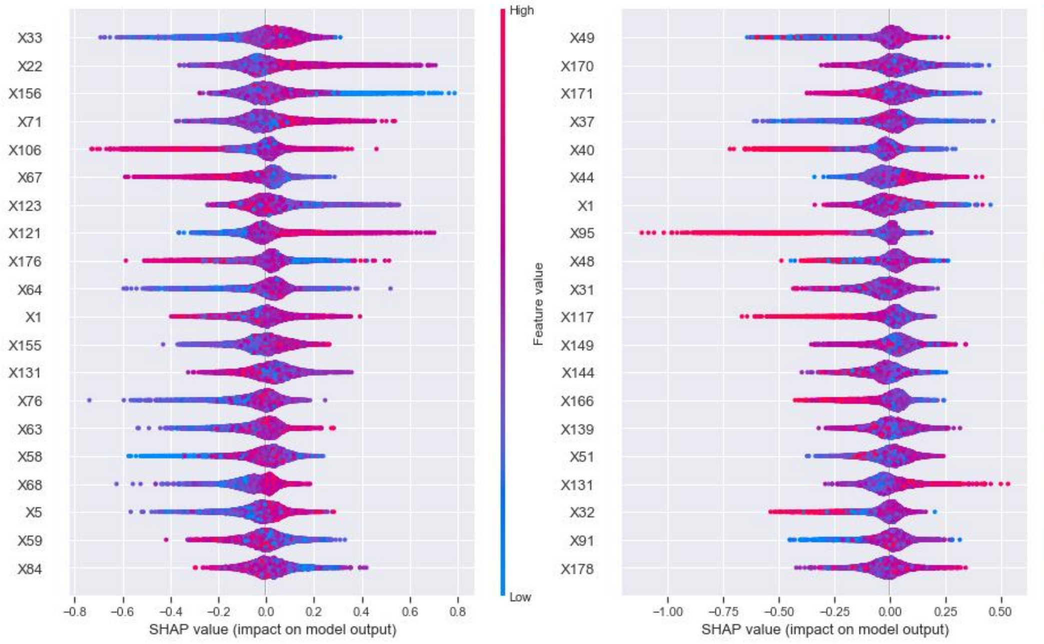


Fig. 26 feature importance of label 2(left) and label 3(right)

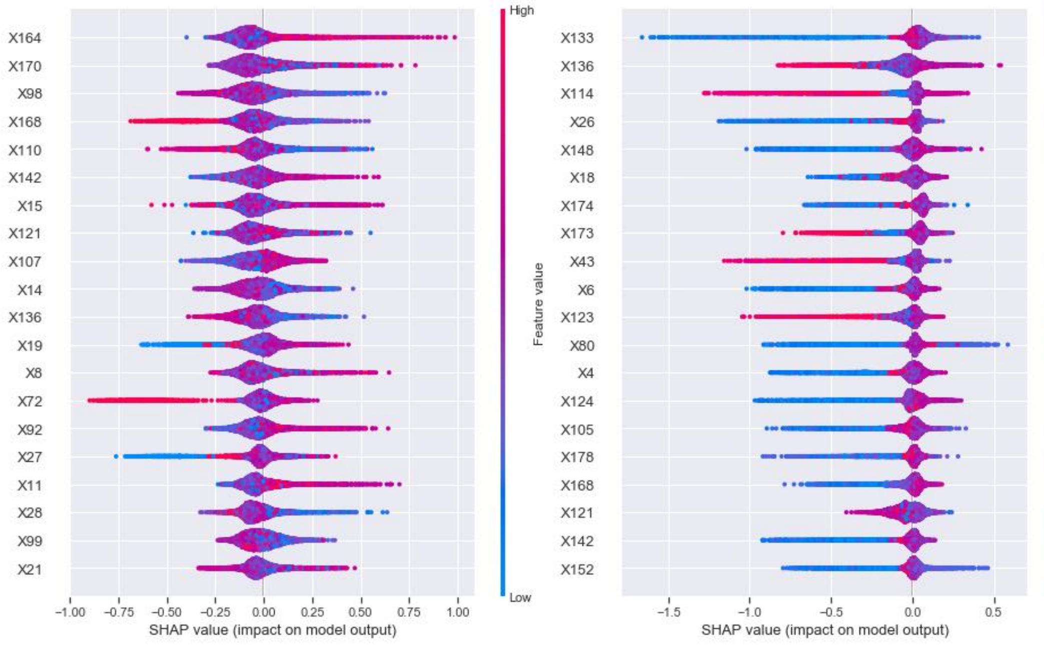


Fig. 27 Feature importance of label 4(left) and label 5(right)

7. Limitation

There are several limitations about this project,

1. Threshold was selected based on the imbalanced sample. If the

unbiased degree of the sample can reflect the true distribution of the whole population, the model can be directly used, but if not, we need to re-select our threshold again.

2. Multi-class classification is carried out by the model fine-tuned by binary classification. The best hyperparameter of multi-class may different with the value of model fine-tune by binary classification.

8. Summary

In this project, the prediction of epileptic seizure was carried out by several base models and ensemble learning models, the results are:

1. EEG of patient with seizure are very different with other normal people., the vibration degree of EEG of patient is very severe.
2. Metrics should be chosen carefully; different metrics will lead to different model.
3. SVM is the best in base machine learning model in the sklearn package tools. And in some range, XGBoost is better than voting classifier with all models .Globally voting with all models is the best model in the ensemble learning method with the weight 4,8,1,2,4,9,11. The best PR-AUC is 0.98808 and positive data f1 score is 0.9419 in best threshold 0.578.
4. For multi-class classification model, voting with all models are still the best model and XGBoost is still the second. The ROC AUC is 0.9338, with the 0.736 f1 score.
5. X1, X2, and X178 is top 3 features for model to predict whether a patient has epileptic seizures. And some values need to be high to lead to seizure while some value need to be low will lead to seizure. For example, the range in 155.14~404.24 of X2 will increase the percentage to get seizure, but X164 with range in -427.28~-176.78 will cause problem.

6. For multi-class classification, X1 X12 and X173 are considered the top 3 important features. For label2, X33, X22 and X156 are the top three ,X49, X170 and X171 are the top 3 for label 3. As for label 4, X164, X170 and X98 are the best when X133, X136 and X114 are considered best for label 5.

9. Future work

1. For real application, we may resample the data to reflect the true population and then use voting classifier to predict labels.
2. Fine-tune the model for multi-class classification to generate best model for predicting.
3. Select the top20 feature importance and use them to predict and evaluate the model performance

10. Reference

[1]. Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. 2006.