Seizure Binary Classification

Using the UCI Epilepsy Dataset

Jerry Yu

The Goal

Predict whether a patient is having a seizure or not

Epilepsy is a disorder of the central nervous system (CNS), affecting about 50 million people worldwide. During a seizure, the brain produces electrical activities that leads to physical symptoms ranging from convulsion, loss of memory, or unconsciousness. In this dataset, the electroencephalogram (EEG) from brain activity is recorded by placing electrodes on the scalp and by measuring the voltage fluctuations between the nodes.

Being able to know whether someone is having a seizure, is the first step to being able to predict whether someone will be having a seizure or not.

Dataset

Context

The dataset contains information information from 500 individuals, each having 4097 data points that cover 23.5 seconds.

Processing

The data points are then split into 23 chunks, each chunk containing 178 data points for each second, and each data point is the value of the EEG recording at a different point in time.

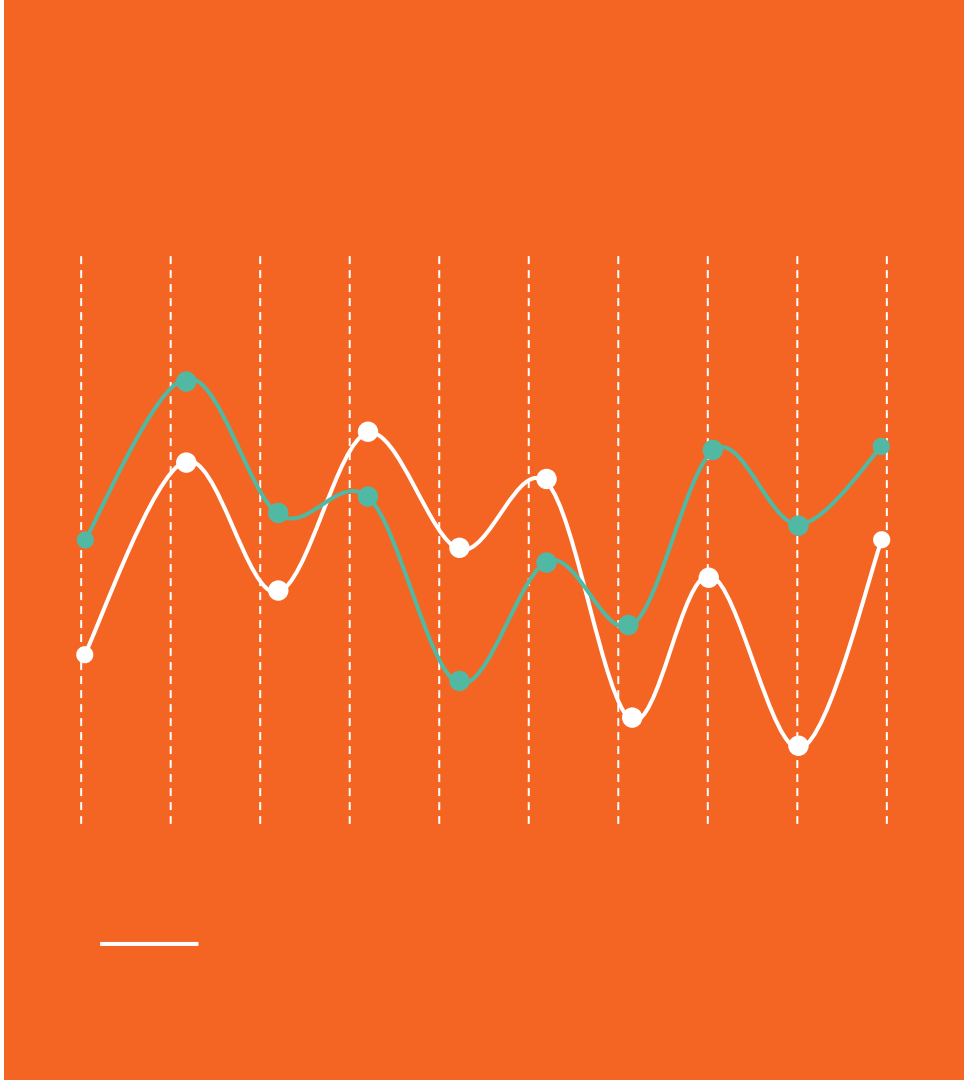
Output

So there are 23 x 500 = 11500 pieces of information (rows), and each row contains 178 data points (columns) for a whole second. The last column contains the binary output.

Dataset: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition

Prevalence

2300 positive samples
/
11500 samples
=
20%

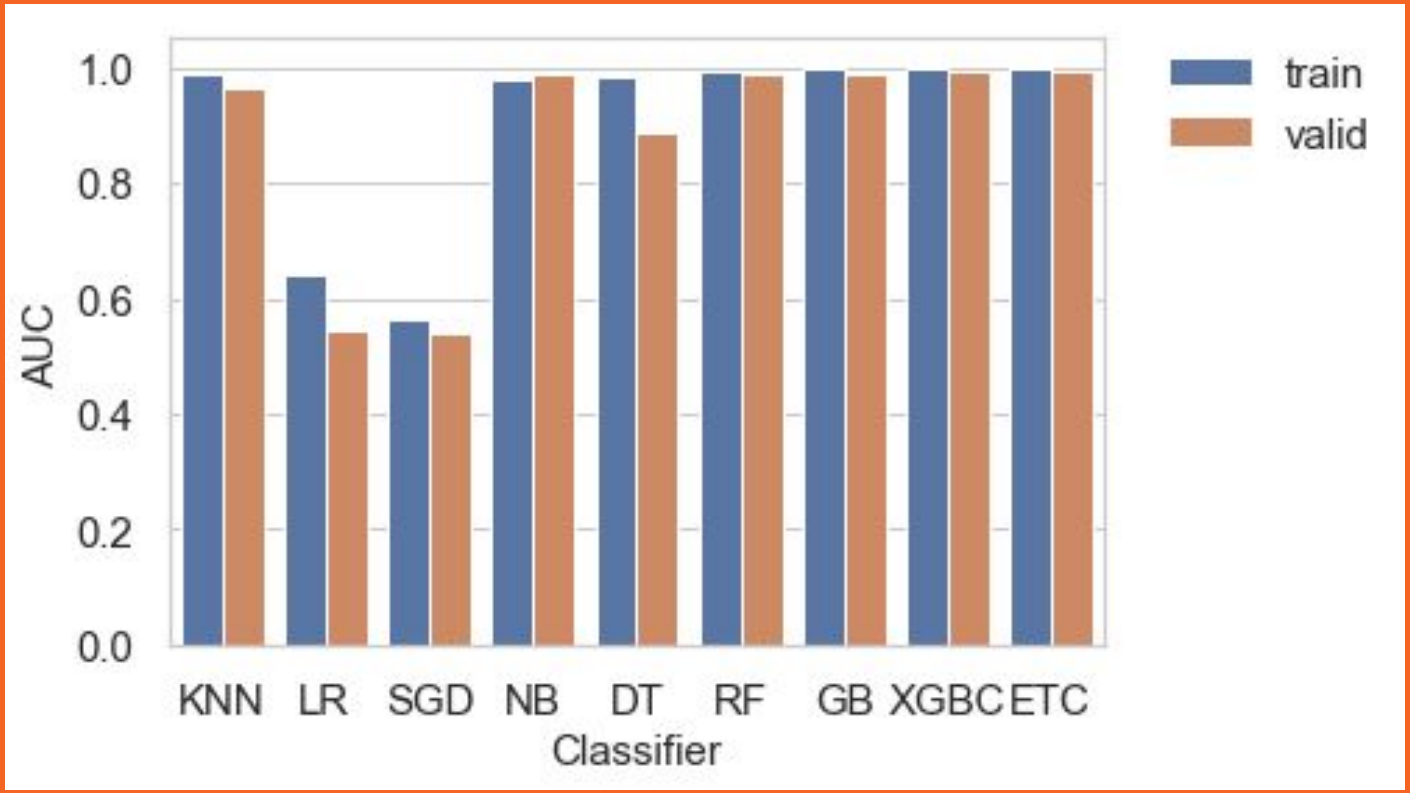


	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X170	X171	X172	X173	X174	X175	X176	X177	X178	OUTPUT_LABEL
0	135	190	229	223	192	125	55	-9	-33	-38	...	-17	-15	-31	-77	-103	-127	-116	-83	-51	0
1	386	382	356	331	320	315	307	272	244	232	...	164	150	146	152	157	156	154	143	129	1
2	-32	-39	-47	-37	-32	-36	-57	-73	-85	-94	...	57	64	48	19	-12	-30	-35	-35	-36	0
3	-105	-101	-96	-92	-89	-95	-102	-100	-87	-79	...	-82	-81	-80	-77	-85	-77	-72	-69	-65	0
4	-9	-65	-98	-102	-78	-48	-16	0	-21	-59	...	4	2	-12	-32	-41	-65	-83	-89	-73	0

5 rows × 179 columns

Feature Engineering

All columns are numerical (EEG reading value), and so no feature engineering is needed.


1.0
0.8
0.6
0.4
0.2
0.0
AUC
KNN
LR
SGD
NB
DT
RF
GB
XGBC
ETC
Classifier
train
valid

XGBoost Classifier

Training AUC = 0.999

Validation AUC = 0.992

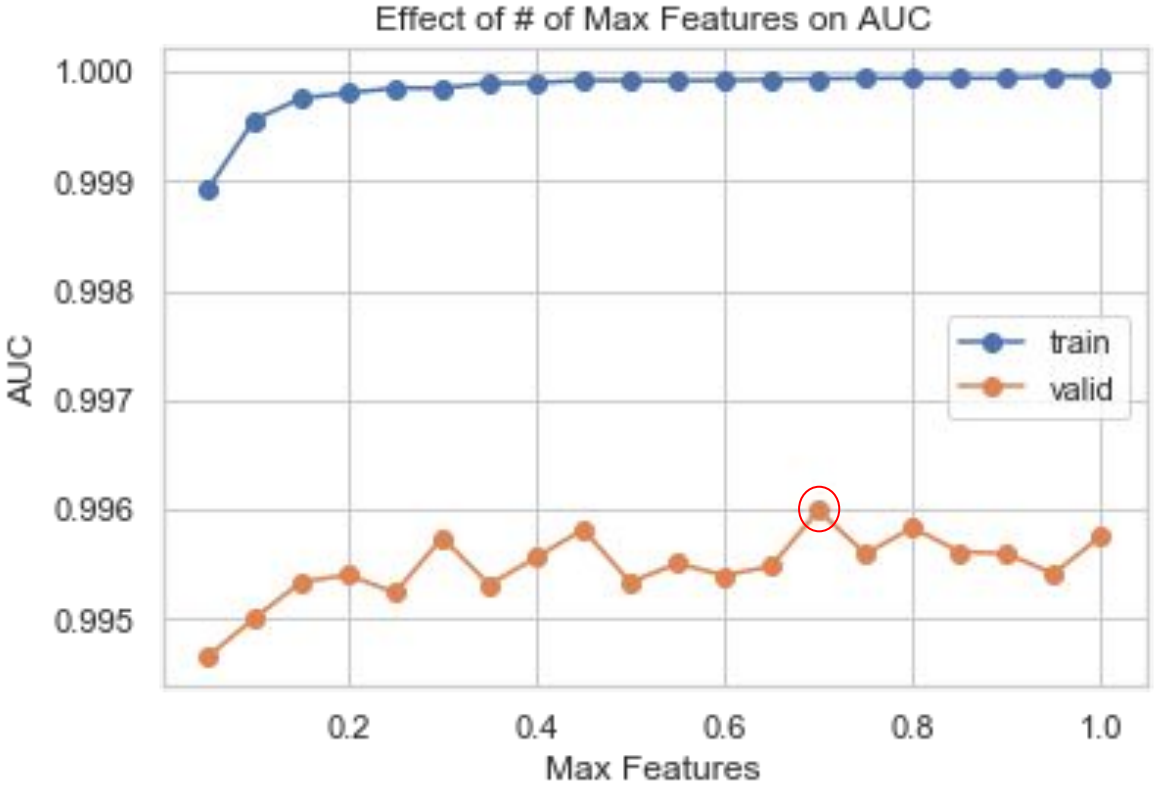
- Trees have a varying number of nodes
- Leaf weights calculated with less evidence is penalized more heavily
- Newton boosting provides a more direct route to the local minima than gradient boosting
- Extra randomization parameter is used to reduce correlation between trees
- Regularized model to prevent over-fitting
- Parallel processing of tree creation

ExtraTrees Classifier

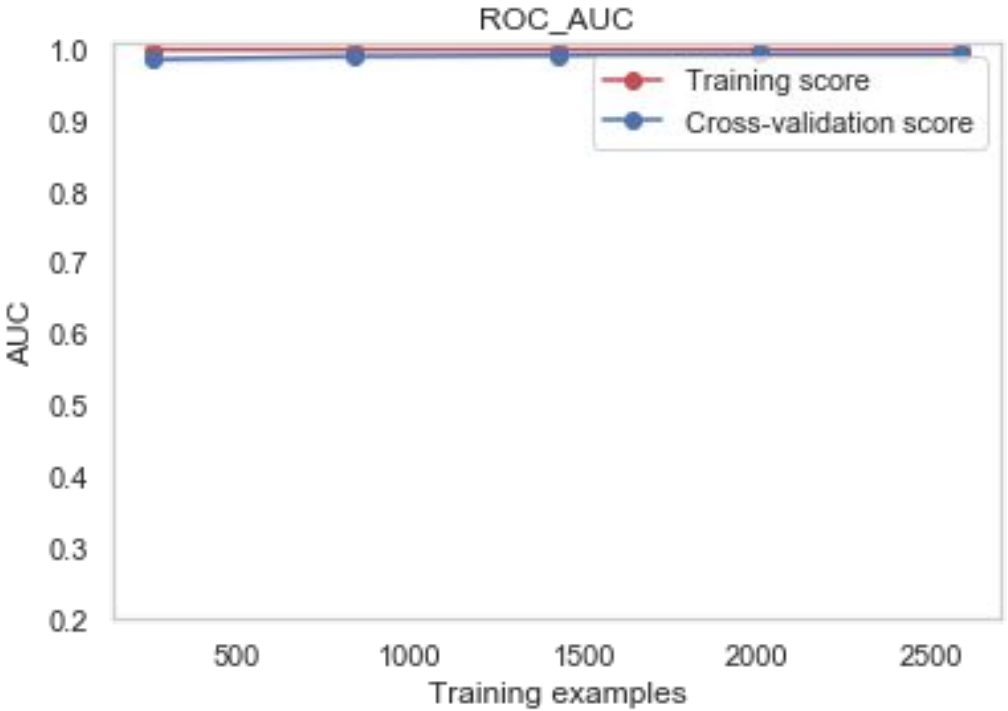
Training AUC = 1.0

Validation AUC = 0.995

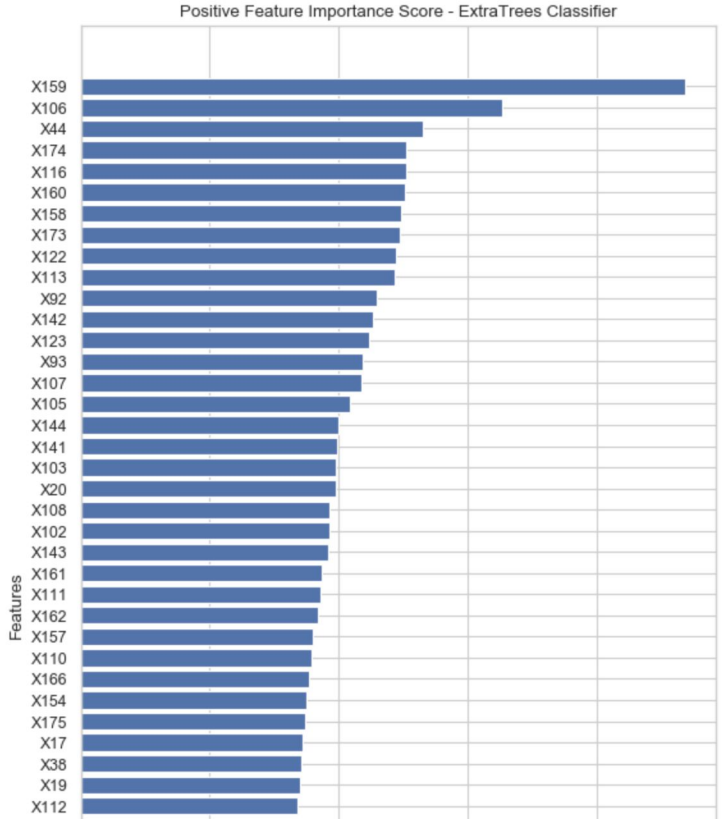
- When nodes are split, the entire dataset is used rather than bootstrapped samples
- Node splits are chosen randomly, rather than testing all possible splits
    - Computationally faster/cheaper
- Out of all tree methods, ExtraTrees often have the best bias/variance tradeoff
- ExtraTrees perform worse when there are noisy or when your data has a high dimensionality



Optimal Max Features: 0.70



Learning Curve



Feature Importance

Genetic Programming

Hyperparameter Optimization
w/
tpot & dask

```
'xgboost.XGBClassifier': {
    'n_estimators': [100],
    'max_depth': range(1, 11),
    'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.],
    'subsample': np.arange(0.05, 1.01, 0.05),
    'min_child_weight': range(1, 21),
    'nthread': [1]
},
'sklearn.ensemble.ExtraTreesClassifier': {
    'n_estimators': [100],
    'criterion': ["gini", "entropy"],
    'max_features': np.arange(0.05, 1.01, 0.05),
    'min_samples_split': range(2, 21),
    'min_samples_leaf': range(1, 21),
    'bootstrap': [True, False]
}
```

- Validation AUC improvement of 0.002 for ExtraTrees Classifier

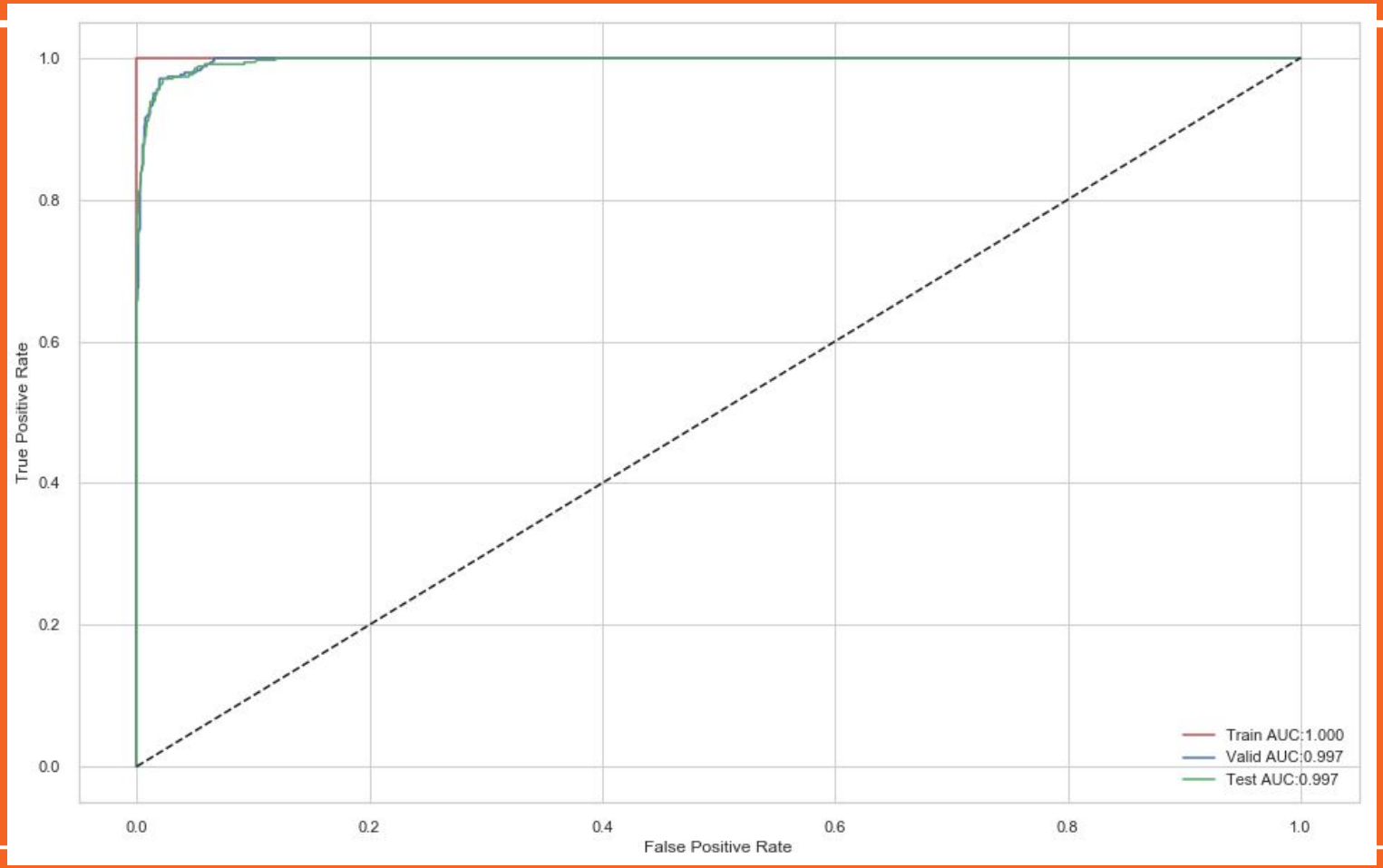
Model Evaluation

Test Set Metrics

```
Training:
AUC:1.000
accuracy:1.000
recall:1.000
precision:1.000
specificity:1.000
prevalence:0.500

Validation:
AUC:0.997
accuracy:0.965
recall:0.977
precision:0.866
specificity:0.962
prevalence:0.199

Test:
AUC:0.997
accuracy:0.959
recall:0.974
precision:0.842
specificity:0.955
prevalence:0.197
```


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Train AUC:1.000
Valid AUC:0.997
Test AUC:0.997

- ExtraTrees Classifier is the best performing model
- 4.3x better than randomly guessing
- 97.4% correct in predicting patients with a seizure correctly
- 99.7% AUC

Thank you!