THE UNIVERSITY OF MELBOURNE

Melbourne University AES/MathWorks/NIH Seizure Prediction

kaggle.com

Team:

nullset

Irina Ivanenko
Oleg Panichev

THE UNIVERSITY OF MELBOURNE

Completed • $20,000 • 478 teams

Melbourne University AES/MathWorks/NIH Seizure Prediction

Fri 2 Sep 2016 – Thu 1 Dec 2016 (13 days ago)

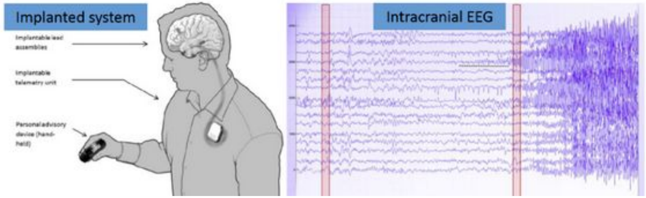
Private Leaderboard

1. Not-so-random-anymore
2. Areté Associates
3. GarethJones
4. QingnanTang
5. nullset
6. tralala boum boum pouêt pouêt
7. Medrr
8. michaln

Predict seizures in long-term human intracranial EEG recordings

Epilepsy afflicts nearly 1% of the world's population, and is characterized by the occurrence of spontaneous seizures. For many patients, anticonvulsant medications can be given at sufficiently high doses to prevent seizures, but patients frequently suffer side effects. For 20-40% of patients with epilepsy, medications are not effective. Even after surgical removal of epilepsy, many patients continue to experience spontaneous seizures. Despite the fact that seizures occur infrequently, patients with epilepsy experience persistent anxiety due to the possibility of a seizure occurring.

Seizure forecasting systems have the potential to help patients with epilepsy lead more normal lives. In order for electrical brain activity (EEG) based seizure forecasting systems to work effectively, computational algorithms must reliably identify periods of increased probability of seizure occurrence. If these seizure-permissive brain states can be identified, devices designed to warn patients of impeding seizures would be possible. Patients could avoid potentially dangerous activities like driving or swimming, and medications could be administered only when needed to prevent impending seizures, reducing overall side effects.



Dashboard

Public Leaderboard - Melbourne Prediction

This leaderboard is calculated on approximately 30% of the test data.
The final results will be based on the other 70%, so the final standings may be different.

#	Δ1w	Team Name * in the money	Score
1	↑16	DataSpring *	0.85457
2	↑1	Not-so-random-anymore *	0.84749
3	↓2	Komaki *	0.84443
4	↑51	Ehsan	0.83372
5	↑11	fugusuki	0.83306
6	↑3	Joseph Chui	0.82696
7	↓5	LabGOL	0.82659
8	↑23	rmldj	0.82114
9	↓1	Mehdi Pedram	0.82088
10	↓5	Kyle	0.82029
11	↓7	Claudia	0.81937
12	↑7	Medrr	0.81851
13	↑1	Alaa-Sean (UWaterloo)	0.81738
14	↓7	GarethJones	0.81524
15	**↓9**	**nullset**	**0.81423**
16	↑125	RNG	0.81216

Dashboard

Private Leaderboard - Melbourne Prediction

This competition has completed. This leaderboard reflects the final standings.

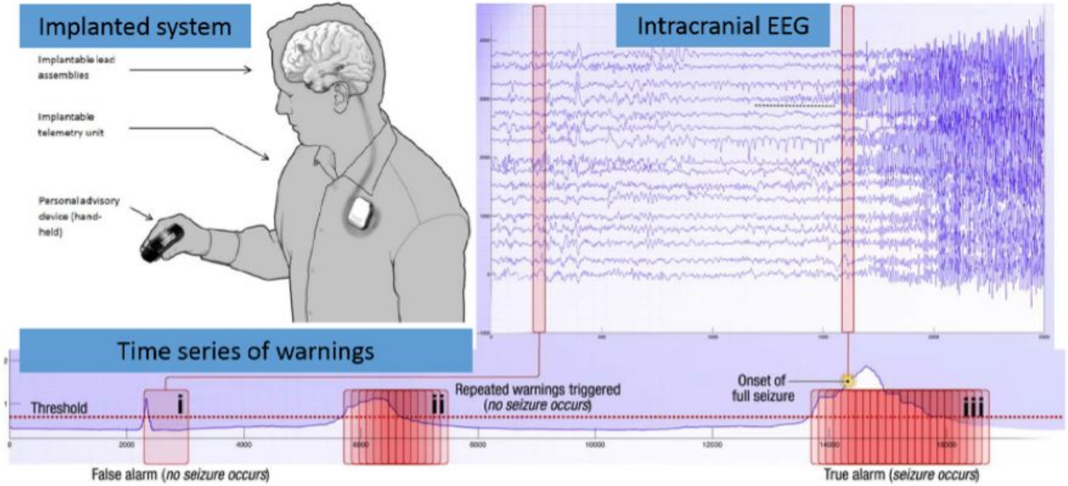
#	Δrank	Team Name ‡ model uploaded * in the money	Score
1	↑1	Not-so-random-anymore ‡ *	0.80701
2	↑35	Areté Associates ‡ *	0.79898
3	↑12	GarethJones ‡ *	0.79652
4	↑23	QingnanTang	0.79458
5	**↑11**	**nullset**	**0.79363**
6	↑14	tralala boum boum pouêt pouêt	0.79197
7	↑7	Medrr	0.79183
8	↑14	michaln	0.79074
9	↓8	DataSpring	0.79053
10	↓5	fugusuki	0.78773
11	↑21	tmunemot	0.78478
12	↓5	Joseph Chui	0.78468
13	↑12	cvanghel	0.78127
14	↓2	krischen	0.77870
15	↑14	QMRSD	0.77778
16	↑5	deepfit	0.77638

Data

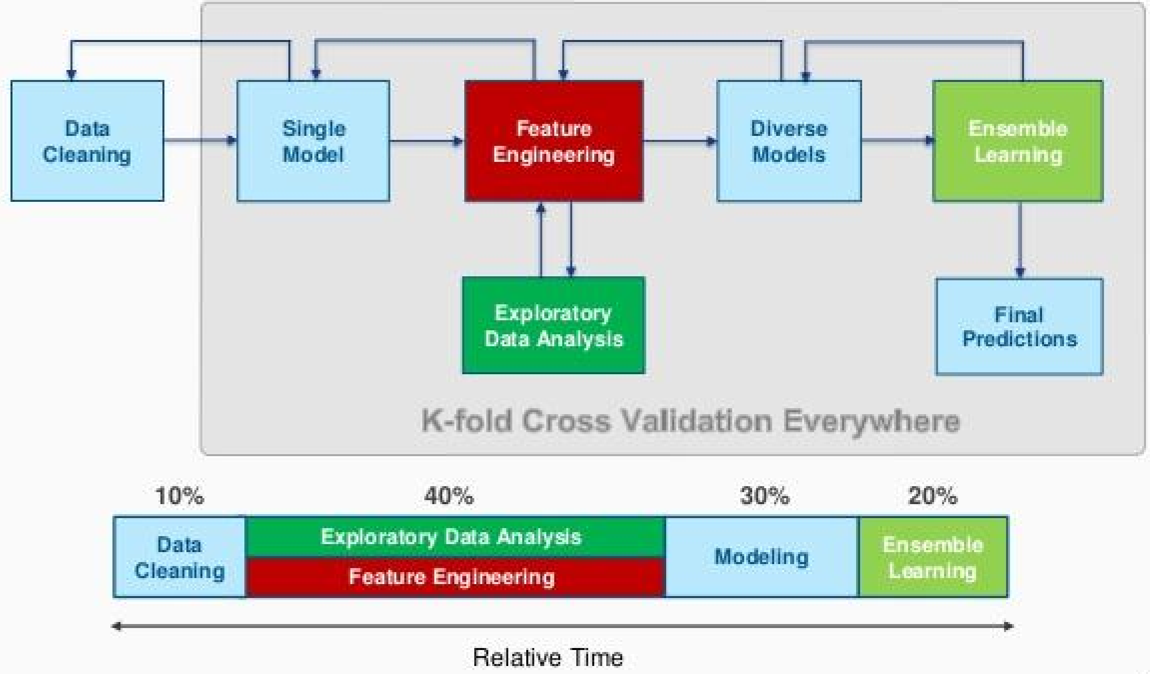
Human brain activity was recorded in the form of intracranial **EEG (iEEG)** which involves electrodes positioned on the surface of the cerebral cortex and the recording of electrical signals with an ambulatory monitoring system.

The challenge is to distinguish between **ten minute long data clips** covering an hour prior to a seizure, and ten minute iEEG clips of interictal activity.

Intracranial EEG (iEEG) data clips **~ 60 Gb** (.mat)



Recommended Data Science Process (IMHO)



Winning data science competitions, presented by Owen Zhang

duration of competition	2 Sep 2016 – 1 Dec 2016
we started	11 Oct 2016
first submission	25 Oct 2016
data leakage and new test set	4 Nov 2016

Software

All data analysis and models were built using Python. Libraries used: scikit-learn, pandas, xgboost.

Preprocessing

The signal from each file was divided on epochs 30 seconds length without any filtration. From each epoch features were extracted. We have tried also 15 and 60 seconds epoch length but the results were worse.

Feature extraction

We tried many features in different combinations during this competition, but not all of them were used in final models. **Feature sets** we've tried:

1. Deep's kernel for features extraction.
2. Tony Reina's kernel for features extraction.
3. Correlation between all channels (120 features).
4. Correlation between spectras of all channels (120 features).
5. Spectral features version 1: total energy (sum of all elements in range 0-30 Hz), energy in delta (0-3 Hz), theta (3-8 Hz), alpha (8-14 Hz) and beta (14-30 Hz) bands, energy in delta, theta, alpha and beta bands divided by total energy, ratios between energies of all bands.
6. Spectral features version 2: the same as Spectral features set 1 plus low and high gamma band were used in calculation of total energy, energy in bands and ratios between energies in bands. In addition, mean energy in bands was extracted.
7. Spectral features version 3: power spectral density was calculated for the whole epoch. Then it was divided on 1 Hz ranges and in each range energy was calculated (30 features).

Fitting and cross-validation

Dividing signals on epochs allowed to increase training dataset size, so total number of observations *No* was equal to

$$No = Nf * Ne,$$

where *Nf* - number of 10-minute signals, *Ne* - number of epochs per one 10-minute signal.

For cross-validation stratified K-folds with 6 folds was used. It was extremely important to use K-fold without shuffling the data, otherwise the leakage is very high and cross-validation performance estimations are much higher. The leakage during shuffling was present because two neighboring epochs with very similar parameters were often present both in train and test sets.

Each model predicted probability of epoch belongs to *preictal* class. The final probability for 10-minute signal was calculated as mean of all probabilities for epochs in this signal.

We tried both patient-specific and non-patient-specific approaches on the same model but performance was higher when patient-specific approach was used.

Models

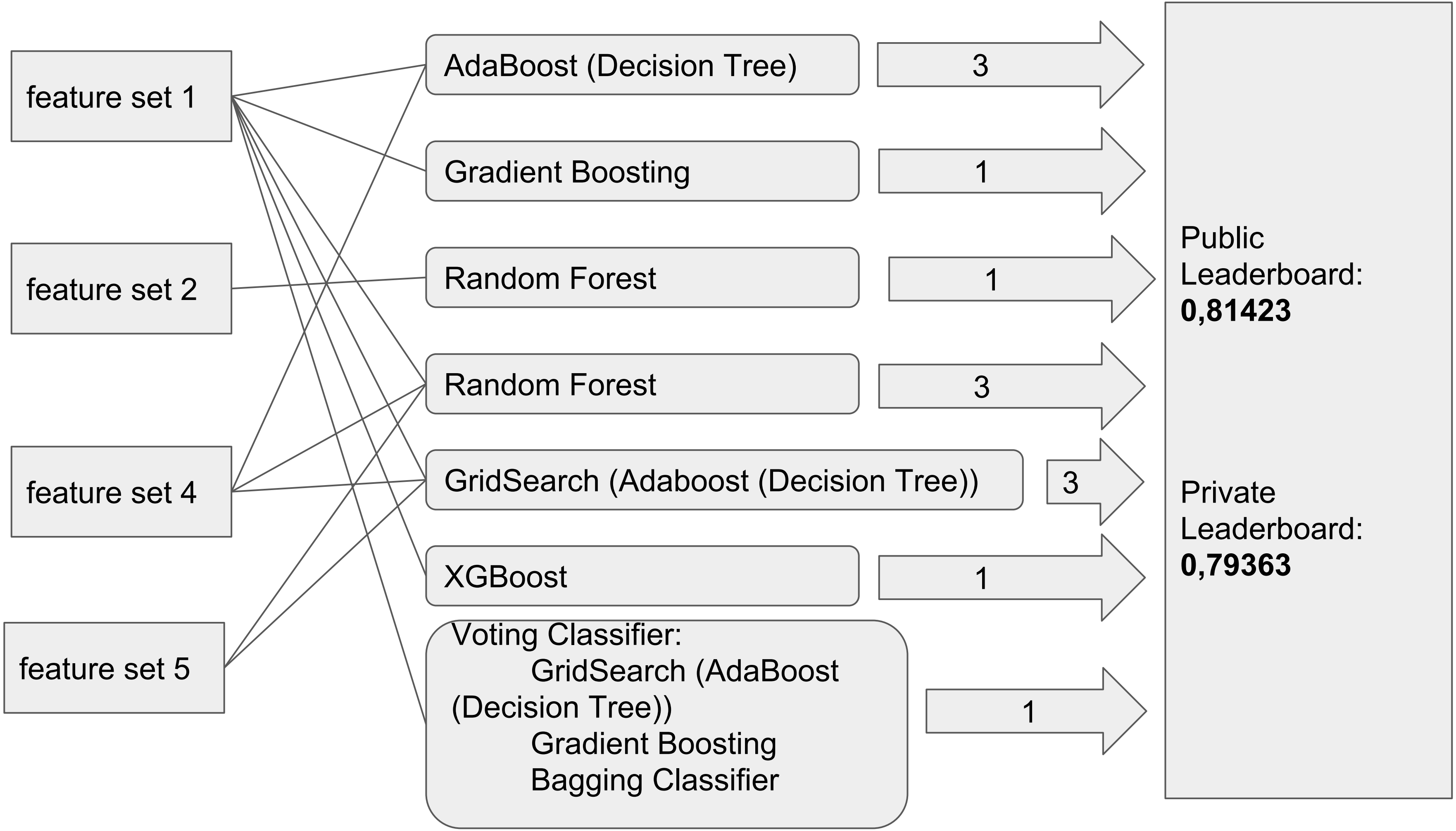
The final solution was an ensemble of best performing models (the first one is the best performing and the last one - is the worst):

1. AdaBoost with Decision Tree base estimator with combined feature sets 1, 4 and 5 .
2. Gradient Boosting Classifier with feature set 1.
3. Random Forest Classifier with feature set 2.
4. Random Forest Classifier with combined feature sets 1, 4 and 5.
5. GridSearch for “number of estimators” parameter for AdaBoost with Decision Tree base estimator with combined feature sets 1, 4 and 5.
6. Voting classifier with feature set 1. Voting was performed for 3 classifiers: GridSearch for “number of estimators” parameter for AdaBoost with Decision Tree base estimator; Gradient Boosting Classifier and Bagging Classifier.
7. XGBoost Classifier with feature set 1.

AdaBoost with Decision Tree base estimator with combined feature sets 1, 4 and 5 showed the highest performance among the models.

Final result *P* was calculated as follows:

P = 1/13 * (3*Model 1 + Model 2 + Model 3 + 3*Model 4 + 3*Model 5 + Model 6 + Model 7)


feature set 1
feature set 2
feature set 4
feature set 5
AdaBoost (Decision Tree)
Gradient Boosting
Random Forest
Random Forest
GridSearch (Adaboost (Decision Tree))
XGBoost
Voting Classifier:
GridSearch (AdaBoost (Decision Tree))
Gradient Boosting
Bagging Classifier
3
1
1
3
3
1
1
Public Leaderboard: **0,81423**
Private Leaderboard: **0,79363**

1-st place solution

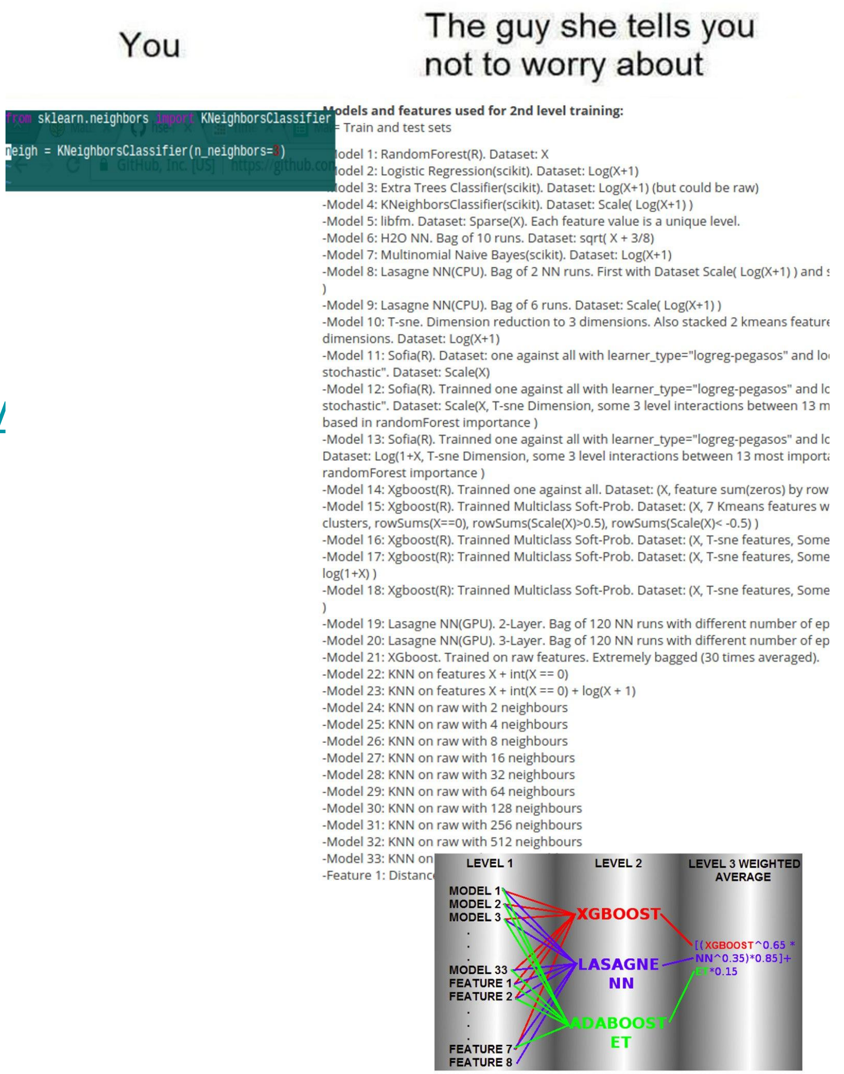
2-nd place on public LB

Aindriú, FengLi, Gilberto Titericz Junior,
Alexandre Barachant

https://www.kaggle.com/c/melbourne-university-seizure-prediction/discussion/26310

Model 1 : XGB, 10 bags, 96 features
Model 2 : XGB, 5 bags, 336 features
Model 3 : XGB, 4 bags, 576 features
Model 4: XGB, 10 bags, 336 features
Model 1: XGB with feature 1
Model 2: KNN with feature1
Model 3: KNN with feature1+feature2
Model 4: Logistic Regression with L2 penalty with feature1+feature2
***Model 1:* All features were used in a bagged XGB classifier (XGB).**
***Model 2*: Linear SVM was trained with top 300 features (SVM)**
***Model 3*: GLM was trained with top 100 features (Glmnet)**

Public: 0.81308 **Private: 0.80701**



Thanks for your attention!