EARTHQUAKE PREDICTION

ISAAC KIM

METIS 2019 FELLOW

CAN WE PREDICT EARTHQUAKES USING MACHINE LEARNING?

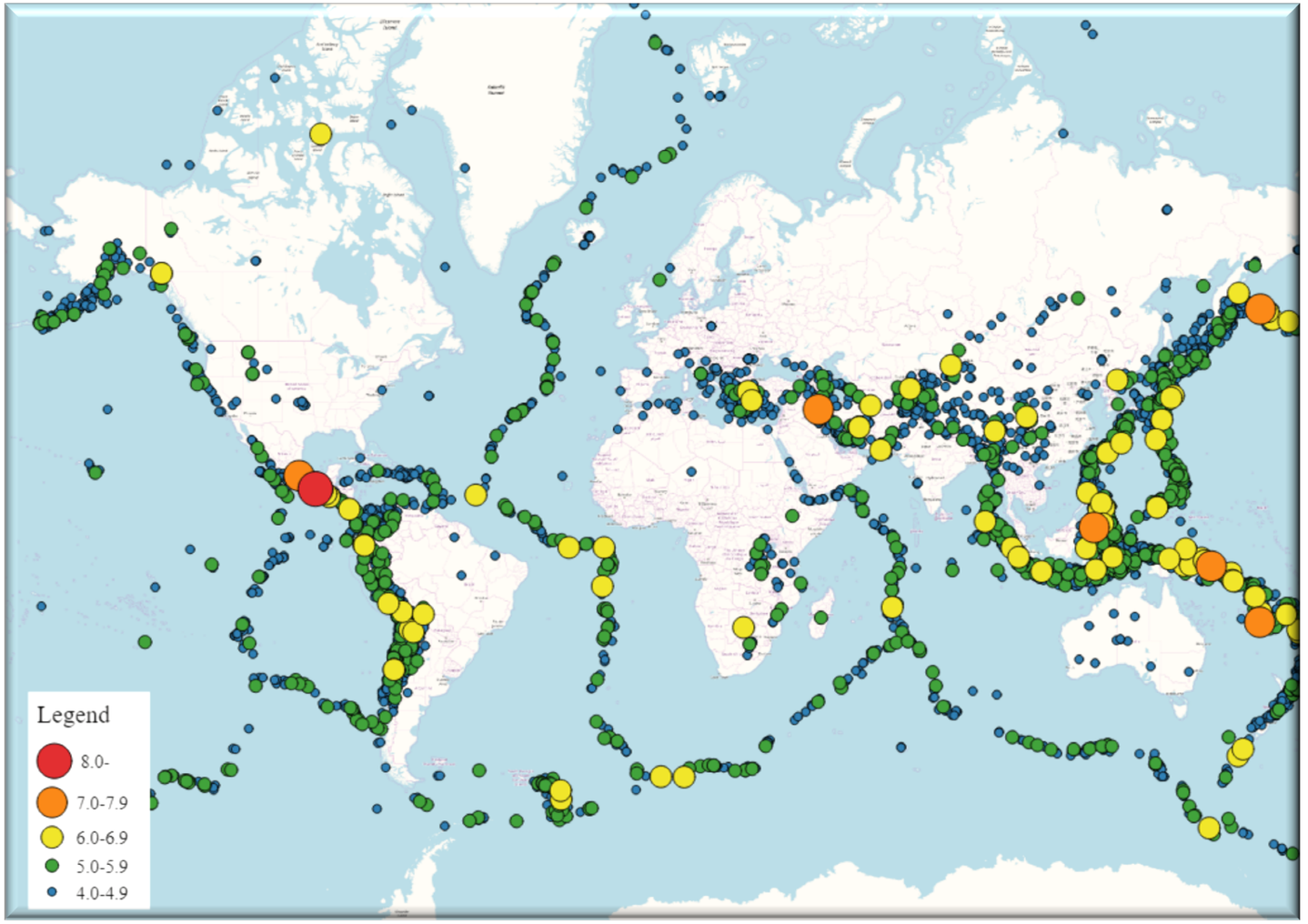
NOPE

BACKGROUND

DIFFICULT

LUCK / PSEUDO-SCIENCE

UNFORESEEN CONSEQUENCES



LABORATORY SETUP

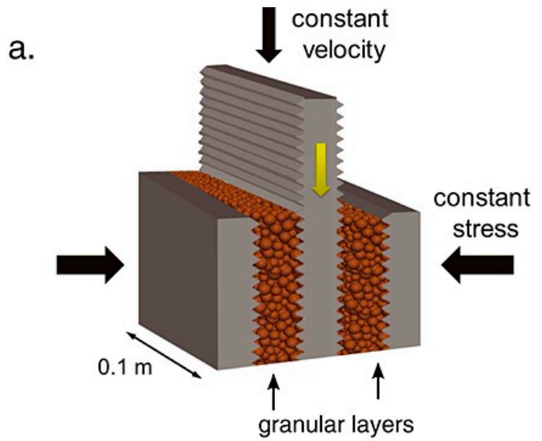
- Two fault gouge layers squeezed
- Vertical layer sheers
- Simulates tectonic activity on fault line

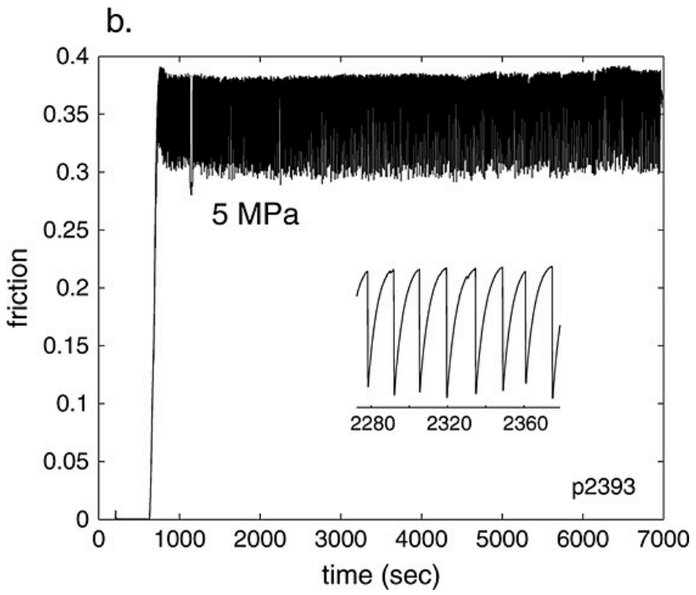
Los Alamos NATIONAL LABORATORY EST.1943

PennState

PURDUE UNIVERSITY

Department of Energy





EXPLORATORY DATA ANALYSIS

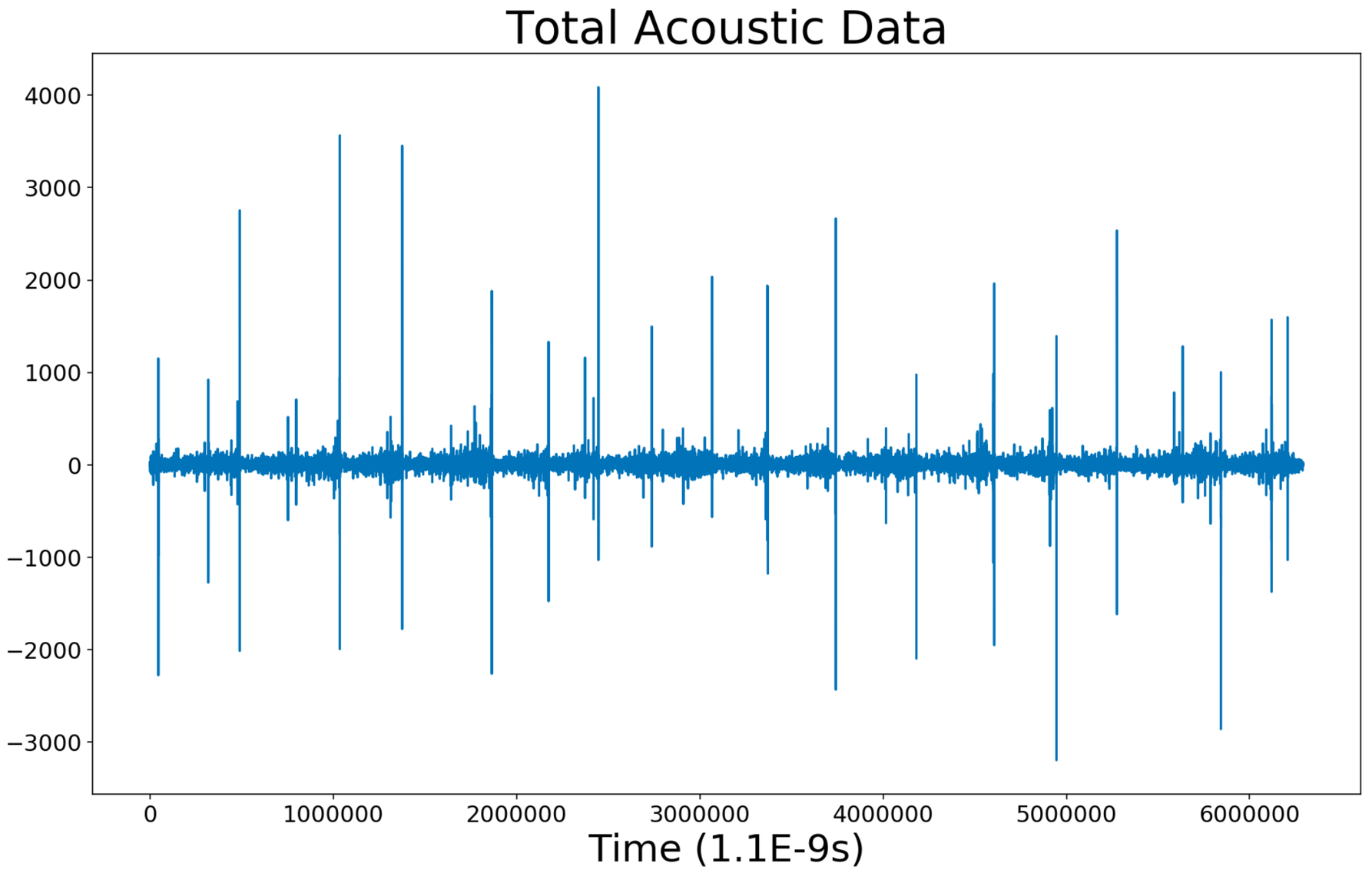
EDA

Training Set

index	acoustic_data	time_to_failure
0	12	1.4691
1	6	1.4691
2	8	1.4691
3	5	1.4691
4	8	1.4691

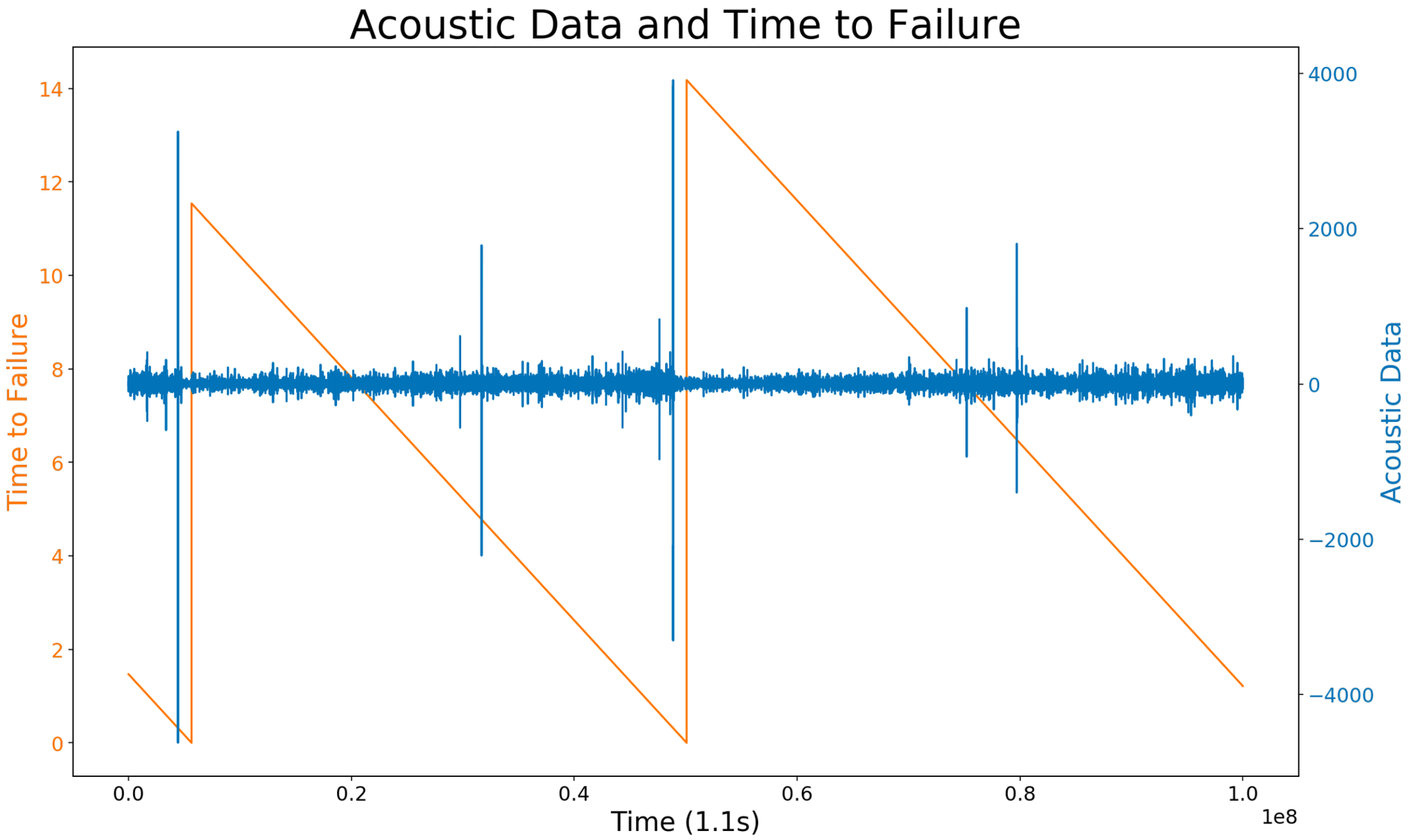
EDA

- 630,000,000 observations
- 1/100 Points sampled here



EDA

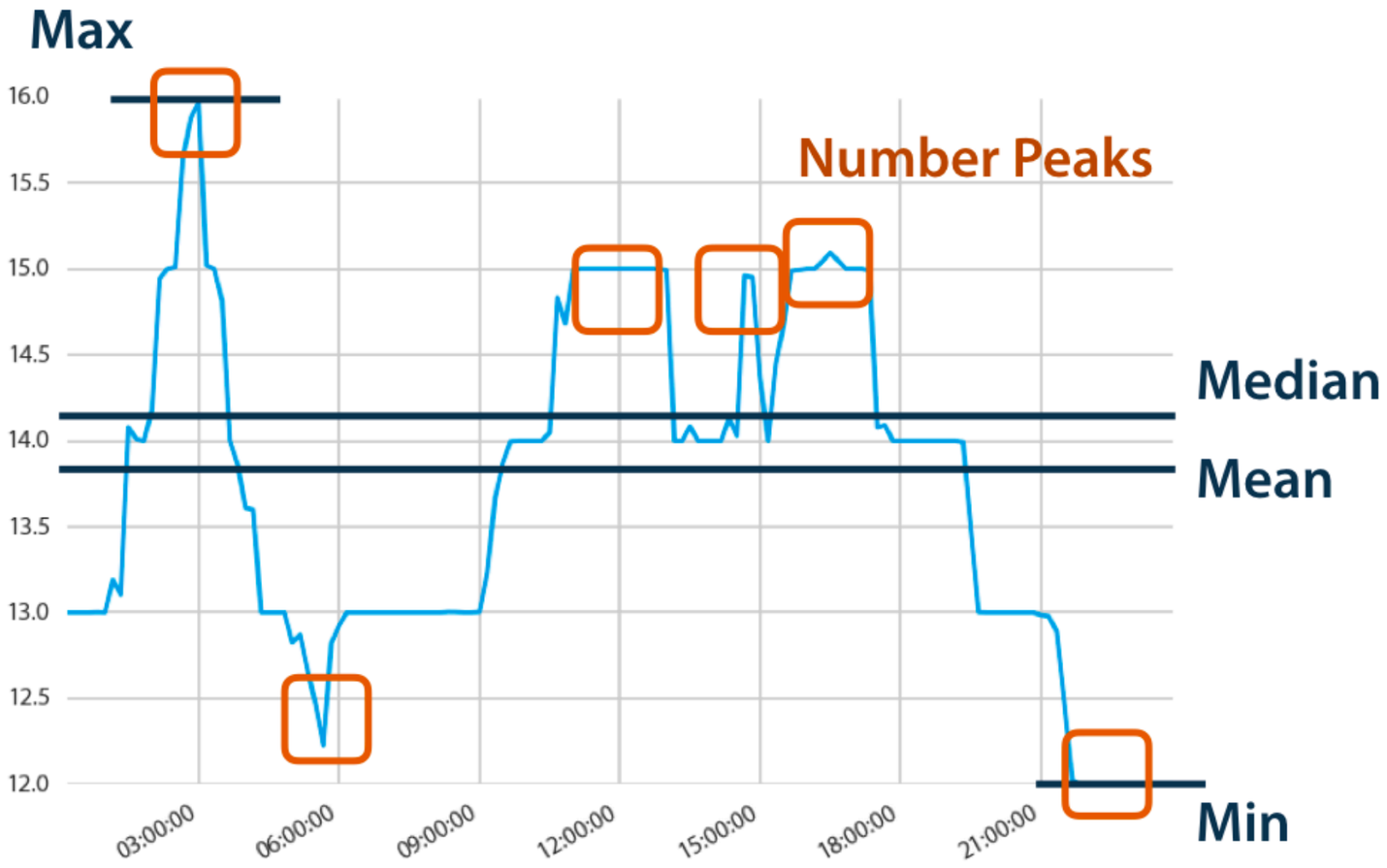
- 16 earthquakes
- Time to failure resets



TSFRESH

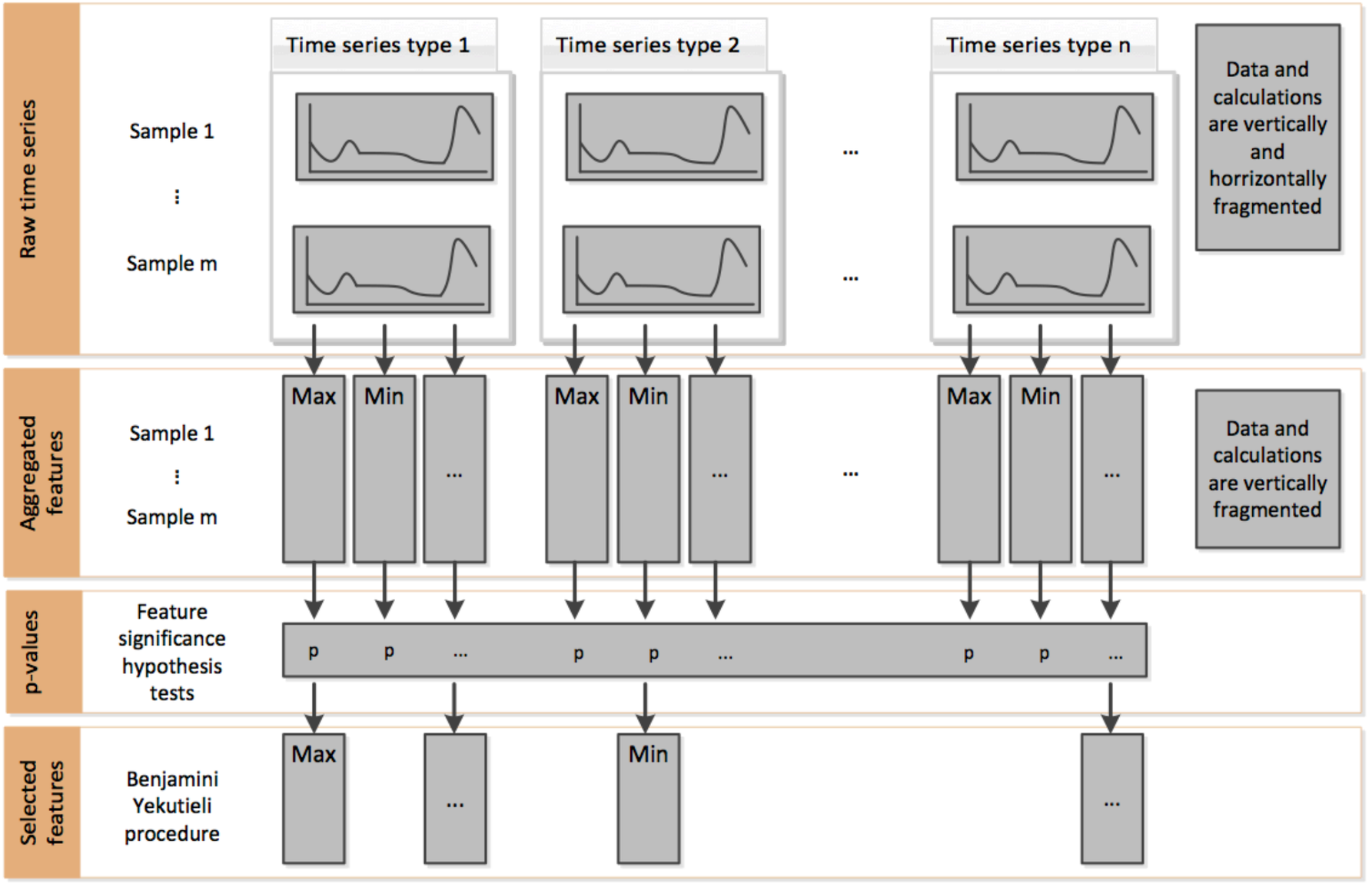
BlueYonder Tech

- Extracts 1200 features from time series
- Selects feature using Benjamini Hochberg Test
- Computationally expensive



TSFRESH

Feature Selection



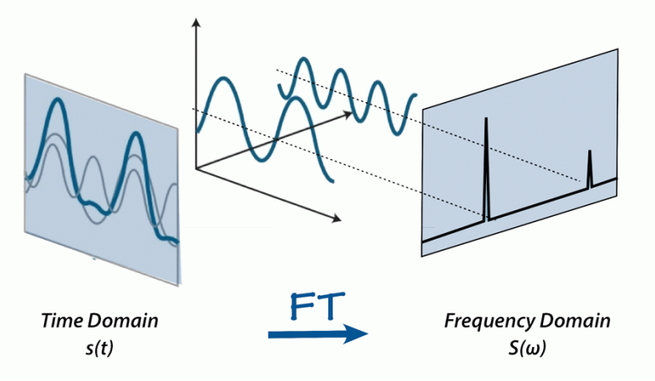
TSFRESH FEATURES

- Autoregressive Model Coefficient

$$X_t = \varphi_0 + \sum_{i=1}^{k} \varphi_i X_{t-i} + \varepsilon_t$$

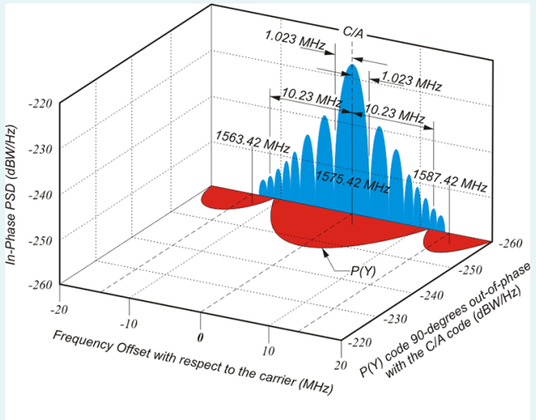
Does current value in time series depend on previous ones?

- Fast Fourier Transform Aggregate Measures



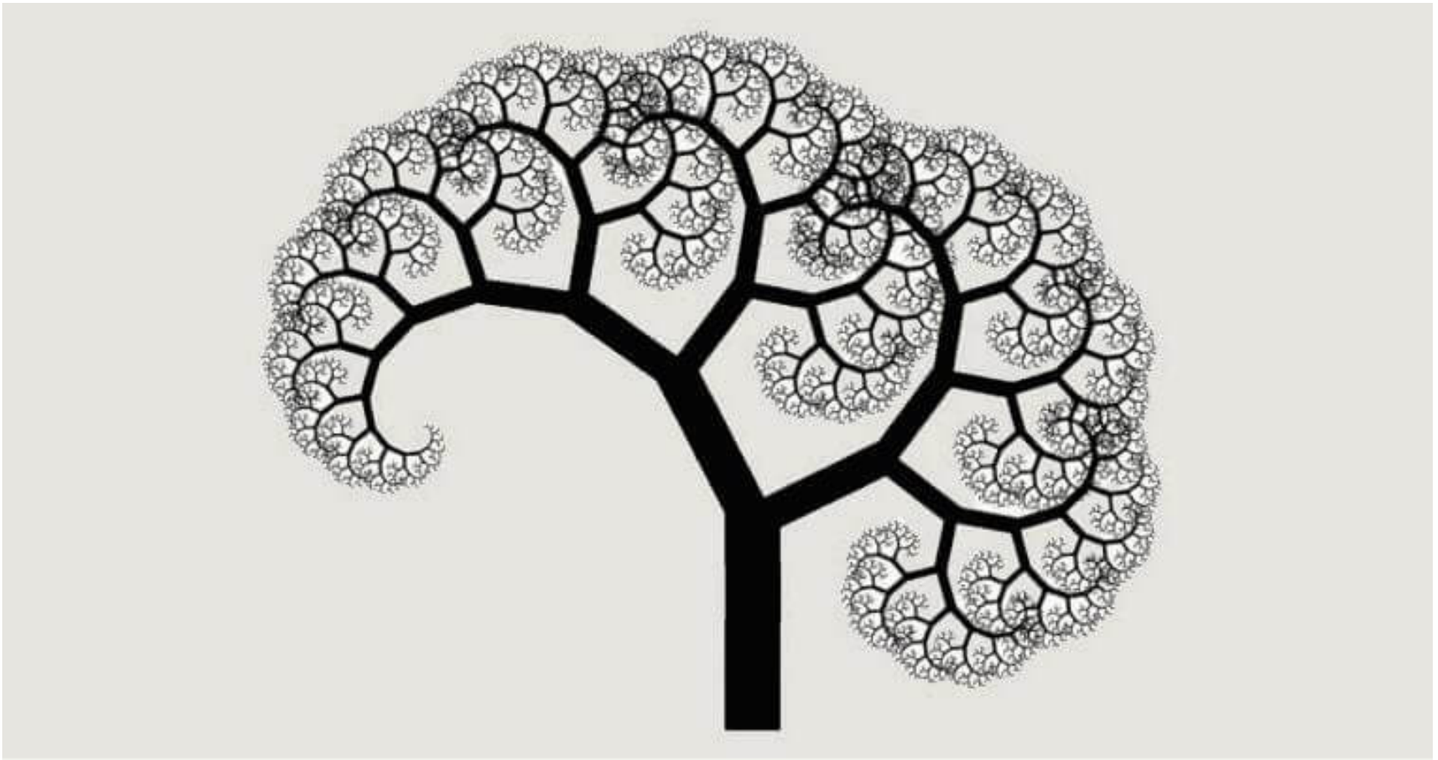
Decompose one signal into several signals.

- Welch Spectral Density



At which frequencies are variations in the window stronger than others?

MODELING



LightGBM, Light Gradient Boosting Machine

LightGBM is a gradient boosting framework that uses tree based learning algorithms.

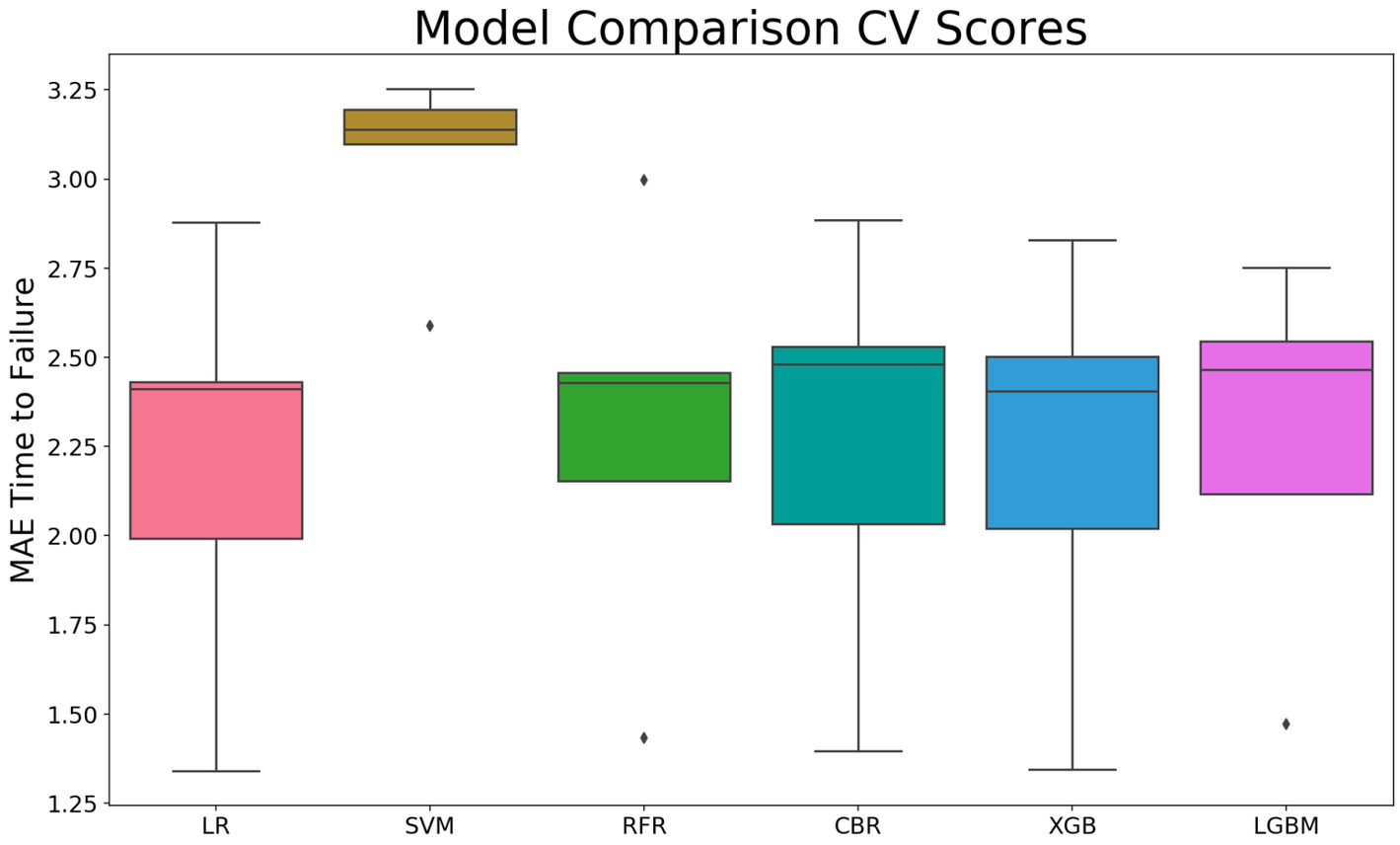
dmlc XGBoost

scikit learn

Yandex CatBoost

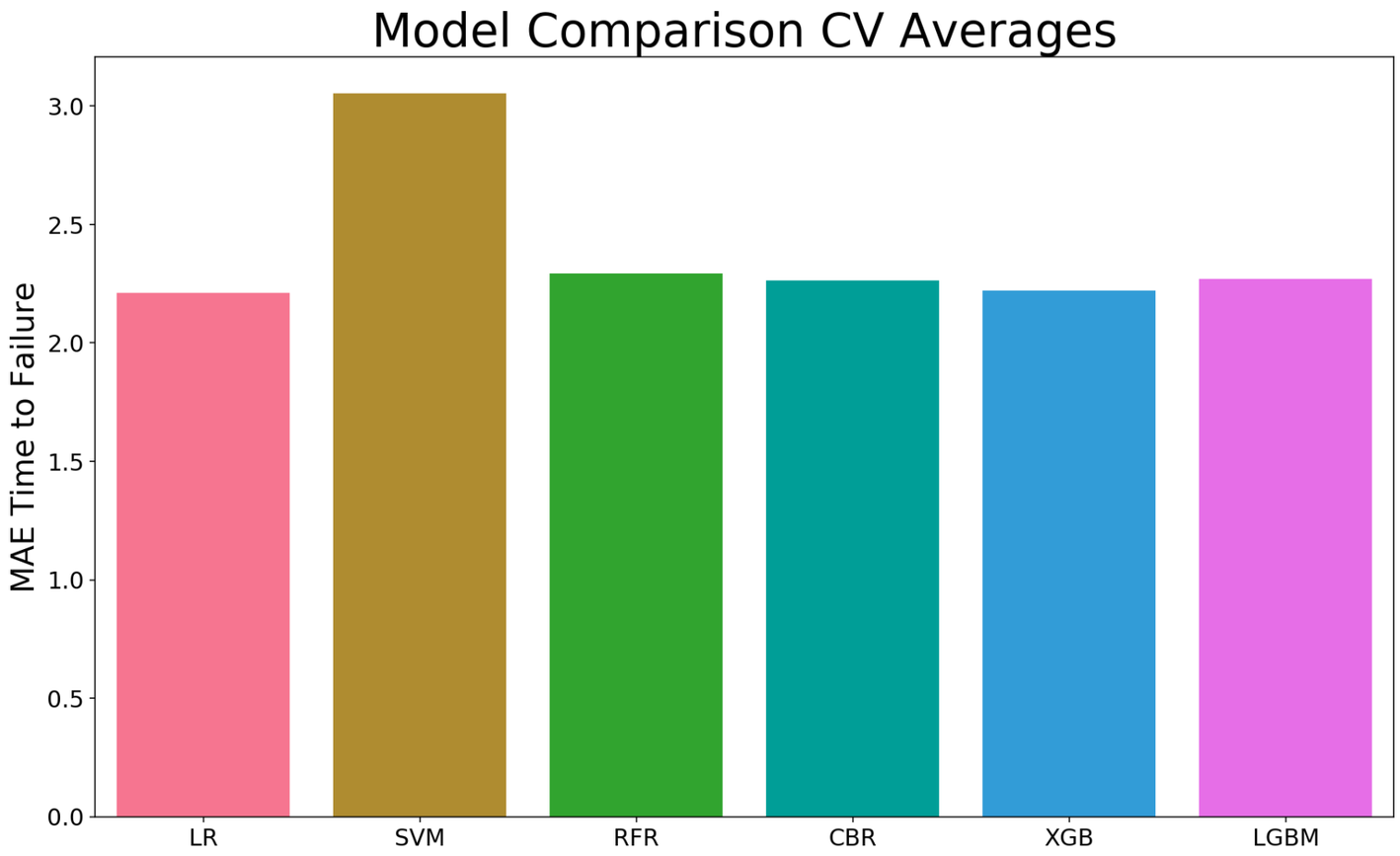
MODELING

- Scoring Metric: Mean Absolute Error

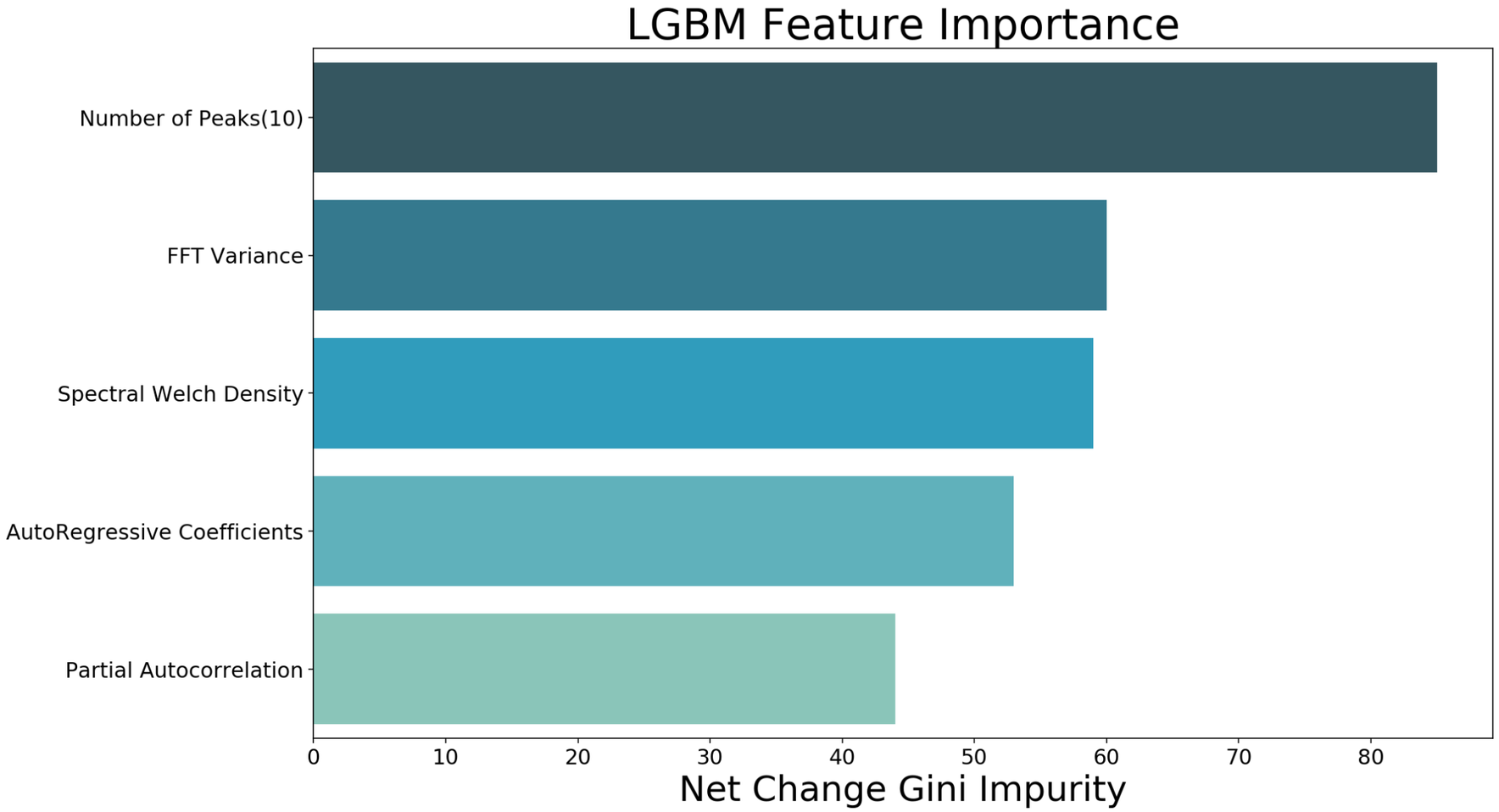


MODELING

- Overfitting major problem with tree regressors
- Key parameters: num_leaves, max_bin, num_iter

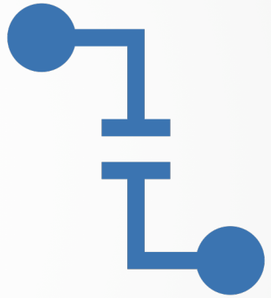


Training Data
MAE Score:
1.99871



RECOMMENDATIONS

Apply these methods to:
mechanical failure in power lines
early detection of solar events.



Protect sensitive systems such as hospitals or nuclear power plants.

THANK YOU

METIS

IsaacNewtonKim

IsaacNewtonKim

towardsdatascience.com/@isaac.kim.d

APPENDICES

Exploratory Data Analysis

Feature Extraction / Selection

Modeling

WORKFLOW

matplotlib NumPy python pandas

ACKNOWLEDGEMENTS

- *Summary and Analysis of The Signal and the Noise: Why so Many Predictions Fail - but Some Dont: Based on the Book by Nate Silver*. Worth Books, 2017.
- Kaggle LANL Competition: Los Alamos National Laboratory, Penn State, Purdue University, Department of Energy

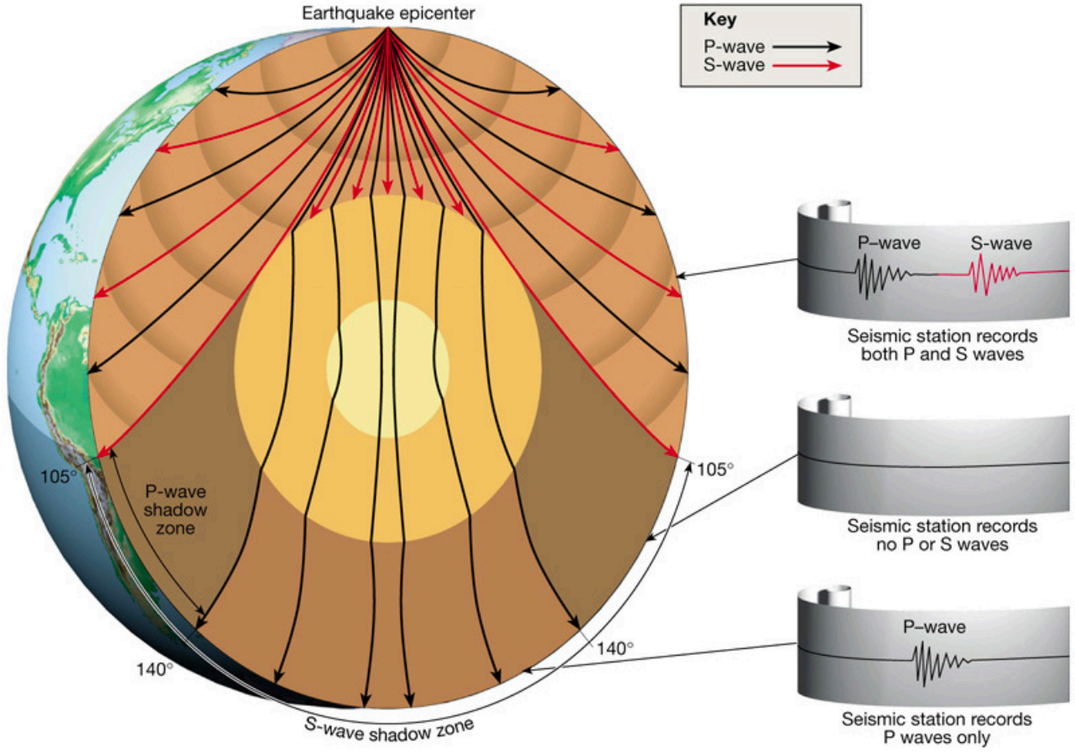
- All the documentation holy crap my eyeballs:
- https://lightgbm.readthedocs.io/en/latest/
- https://tsfresh.readthedocs.io/en/latest/ *Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests*
- https://www.sciencedirect.com/science/article/pii/S0925231218304843?via%3Dihub
- https://xgboost.readthedocs.io/en/latest/

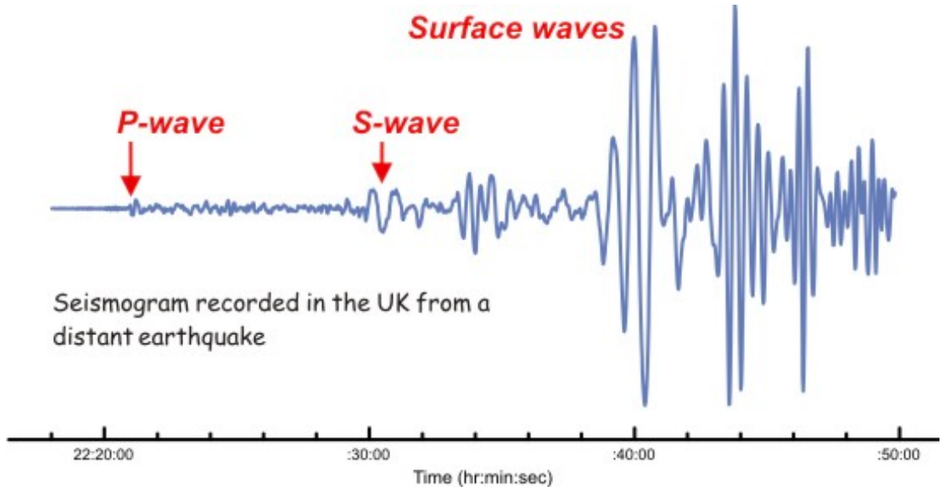
S AND P WAVES

Primary Wave = Compressional

S Wave = Transverse

Surface Wave = Damage





TSFRESH

Here are the extracted features I used.

- The features must be written in a very specific format as a dictionary.
- Each feature can have hyperparameters.

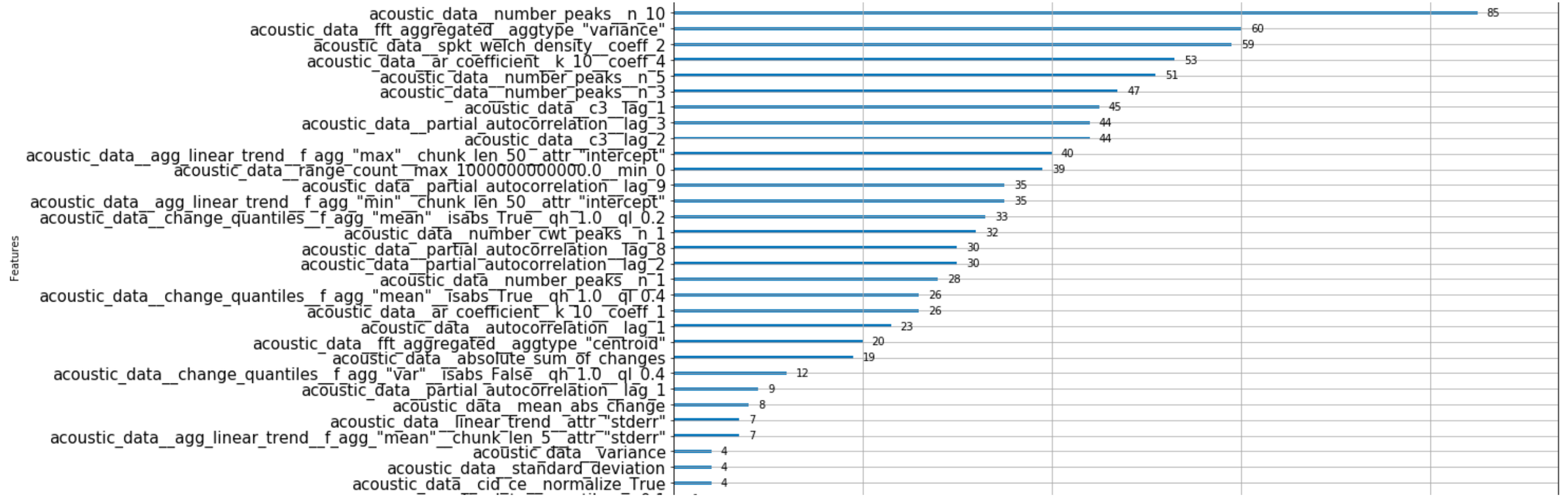
```
smaller2 = {'acoustic_data': {'number_peaks': [{'n': 3}, {'n': 1}, {'n': 5}, {'n': 10}],
  'change_quantiles': [{'f_agg': 'mean', 'isabs': True, 'qh': 1.0, 'ql': 0.2},
   {'f_agg': 'mean', 'isabs': True, 'qh': 1.0, 'ql': 0.4},
   {'f_agg': 'var', 'isabs': False, 'qh': 1.0, 'ql': 0.4}],
  'range_count': [{'max': 1000000000000.0, 'min': 0}],
  'number_cwt_peaks': [{'n': 1}],
  'mean_abs_change': None,
  'absolute_sum_of_changes': None,
  'c3': [{'lag': 1}, {'lag': 2}],
  'ar_coefficient': [{'k': 10, 'coeff': 4}, {'k': 10, 'coeff': 1}],
  'partial_autocorrelation': [{'lag': 2},
   {'lag': 8},
   {'lag': 3},
   {'lag': 9},
   {'lag': 1}],
  'quantile': [{'q': 0.1}],
  'fft_aggregated': [{'aggtype': 'variance'}, {'aggtype': 'centroid'}],
  'spkt_welch_density': [{'coeff': 2}],
  'agg_linear_trend': [{'f_agg': 'min', 'chunk_len': 50, 'attr': 'intercept'},
   {'f_agg': 'mean', 'chunk_len': 5, 'attr': 'stderr'},
   {'f_agg': 'max', 'chunk_len': 50, 'attr': 'intercept'}],
  'variance': None,
  'standard_deviation': None,
  'linear_trend': [{'attr': 'stderr'}],
  'cid_ce': [{'normalize': True}],
  'autocorrelation': [{'lag': 1}]}}
```

LGBM PARAMETERS

```
gridParams = {
    'learning_rate': [0.005, 0.01,.1],
    'n_estimators': [50,100,150],
    'num_leaves': [4,6,7,8], # large num_leaves helps improve accuracy but might lead to over-fitting
    'boosting_type' : ['gbdt', 'dart'], # for better accuracy -> try dart
    'objective' : ['mean_absolute_error'],
    'max_bin':[127,255], # large max_bin helps improve accuracy but might slow down training progress
    'colsample_bytree' : [.4,.6,.8],
    'subsample' : [.4,.6,.7,.8],
    'reg_alpha' : [0,.5,1],
    'reg_lambda' : [0,.5,1],
    }
```

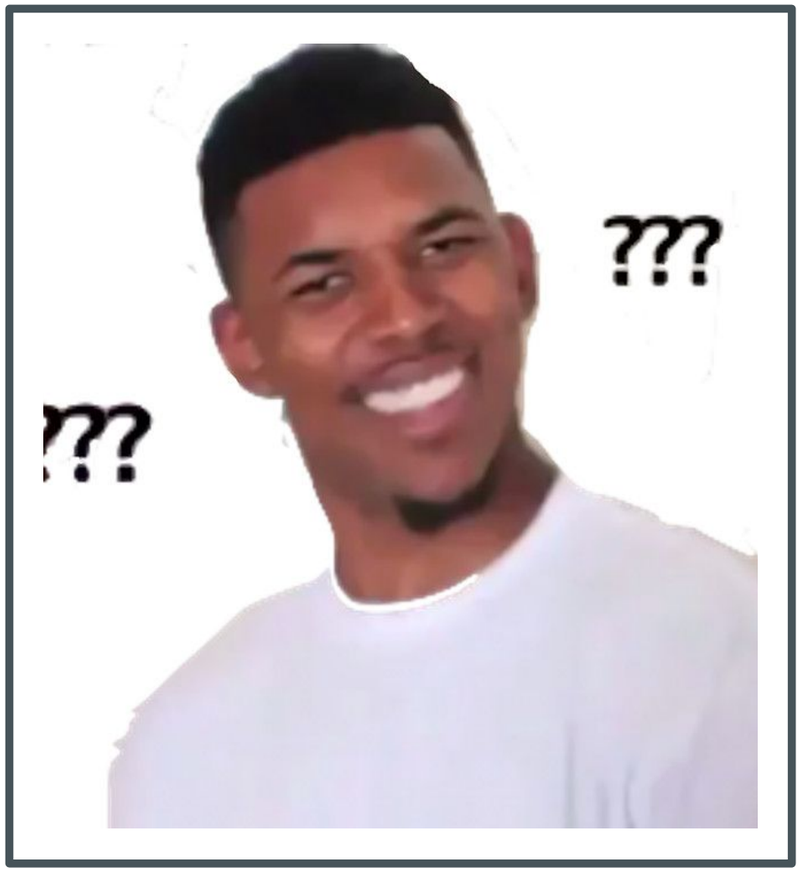
> This was the winner!

```
lgbm = LGBMRegressor(colsample_bytree=0.56,
                     max_bin= 255, n_estimators= 120, num_leaves= 9,
                     objective= 'mean_absolute_error', subsample= 0.7)
```



FEATURE IMPORTANCE

TERMINOLOGY I DON'T FULLY UNDERSTAND



- A fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IDFT). Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. The FFT reduces computation time from $O(N^2)$ to $O(NlogN)$.
- The statistical average of a certain signal or sort of signal (including noise) as analyzed in terms of its frequency content, is called its spectrum.
- Power spectral density function (PSD) shows the strength of the variations(energy) as a function of frequency. In other words, it shows at which frequencies variations are strong and at which frequencies variations are weak.

FAST FOURIER TRANSFORM

$$\sum_{n=0}^{N-1} a_n \, e^{-2\pi i \, n \, k/N} = \sum_{n=0}^{N/2-1} a_{2\,n} \, e^{-2\pi i \, (2\,n) \, k/N}$$

$$+ \sum_{n=0}^{N/2-1} a_{2\,n+1} \, e^{-2\pi i \, (2\,n+1) \, k/N}$$

$$= \sum_{n=0}^{N/2-1} a_n^{\text{even}} \, e^{-2\pi i \, n \, k/(N/2)}$$

$$+ e^{-2\pi i \, k/N} \sum_{n=0}^{N/2-1} a_n^{\text{odd}} \, e^{-2\pi i \, n \, k/(N/2)},$$

CONTEST RANKINGS

19 submissions for isaackim0537 Sort by Most recent

All Successful Selected

Submission and Description	Private Score	Public Score	Use for Final Score
finallgbm.csv 6 hours ago by isaac add submission details	2.47125	1.60837	☐

In the money Gold Silver Bronze

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	▲ 354	The Zoo		+5	2.26589	238	3mo
2	▲ 671	Jun Koda			2.29670	69	3mo
3	▲ 77	Character Ranking			2.29686	96	3mo
4	▲ 259	Reza			2.29749	46	3mo
5	▲ 174	Glory or death!			2.29801	24	3mo
112	▲ 1264	liuyuan0811			2.46961	14	4mo
113	▲ 1627	jiiteecee			2.47055	29	3mo
114	▲ 2344	Devon Bridgeman			2.47100	14	7mo