Kaggle

Stay Alert! The Ford Challenge

Dmytro Kolesnykov

Data understanding

Panel data (also known as longitudinal or cross-sectional time-series data) is a dataset in which the behavior of entities are observed across time. In this case period of around 2 minutes of driving produce sequential data with a unique trial ID.

P1, P2 , ……., P8 physiological data;

E1, E2, ……., E11 environmental data;

V1, V2, ……., V11 vehicular data;

Is_Alert target data : { 0 ; 1 }

Choosing suitable approach

- ❑ Statistical:

GLMM, GEE, VAR, Markov models

- ❑ Signal processing:

Fourier, wavelet power spectrum

- ❑ Neural Networks:

RNN, LSTM

✓Casual Kaggle-style ML:

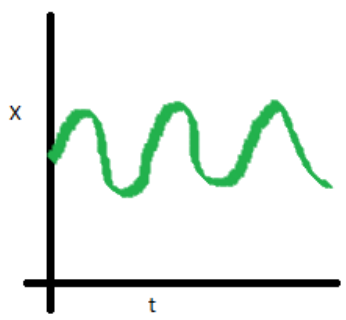
Data preparation

Feature engineering

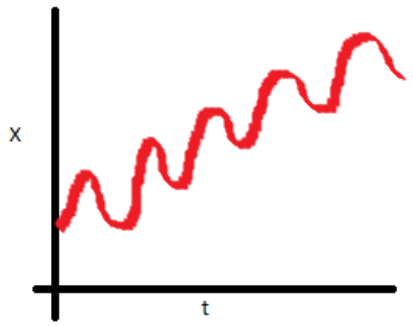
- Get rid of correlated variables and useless features or trials
- Rolling statistics, aggregates
- Lagged values
- Absolute differences

Transformations

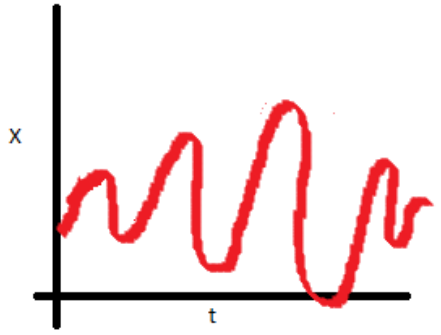
- Detrending
- Square-root
- Log
- Box-Cox



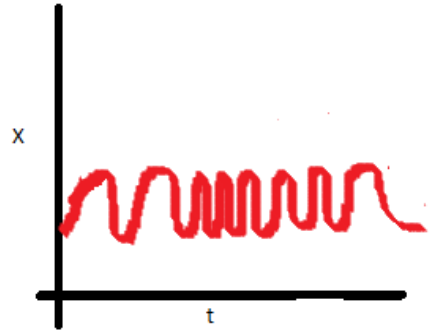
Stationary series



Non-Stationary series

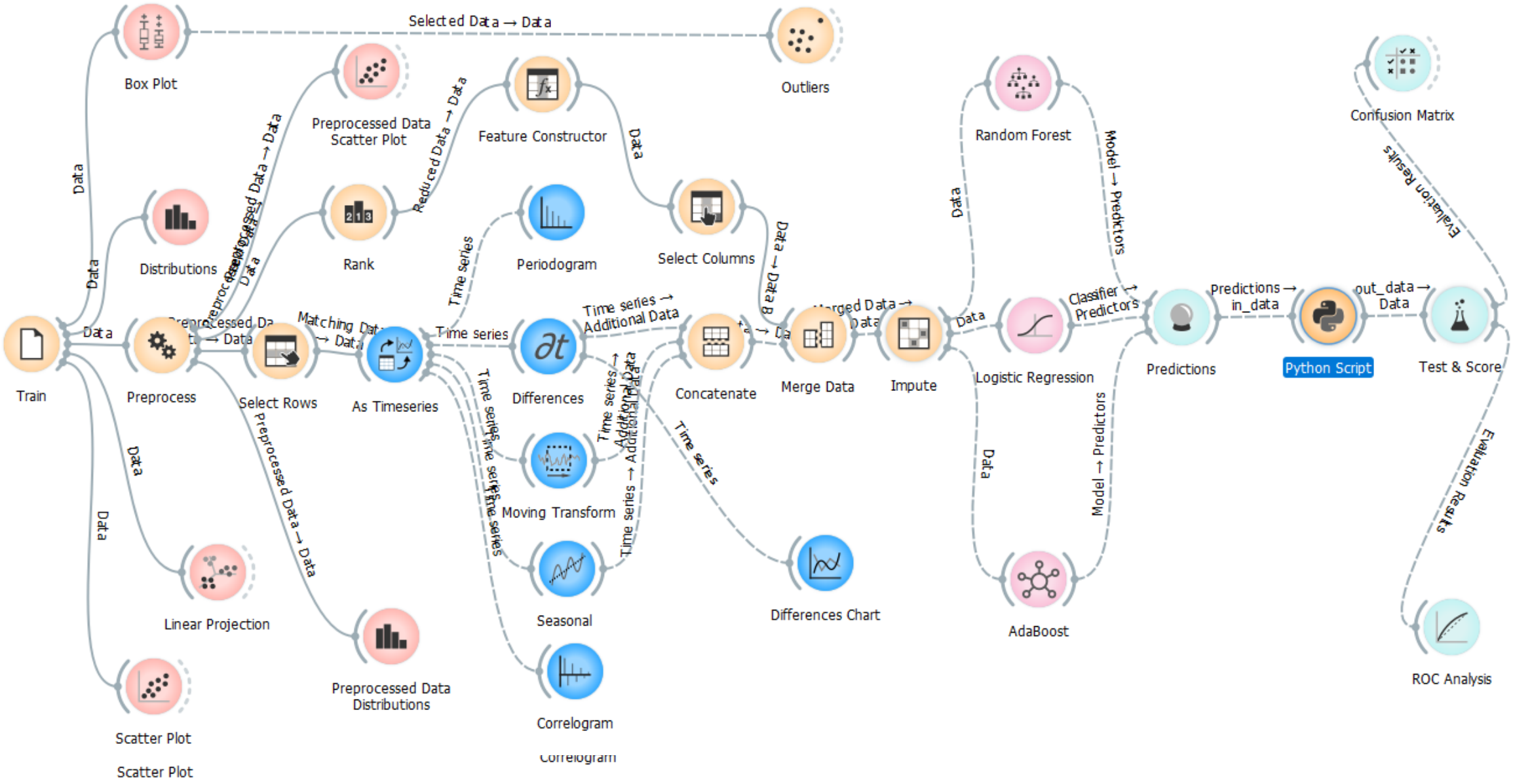


Non-Stationary series



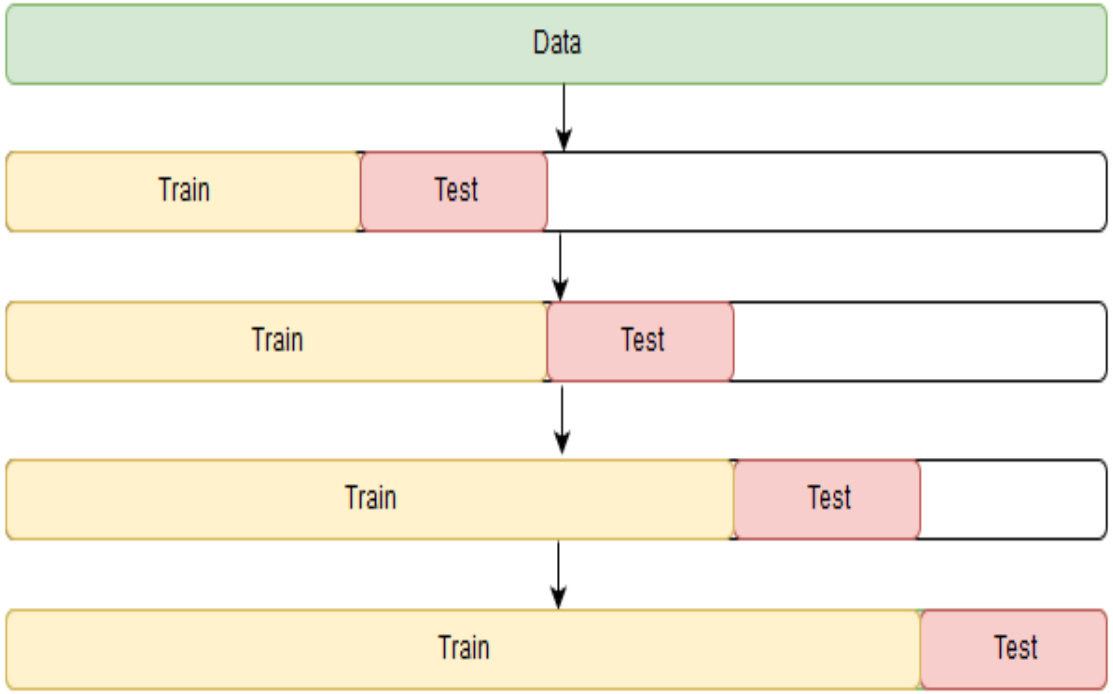
Non-Stationary series

Modeling pipeline

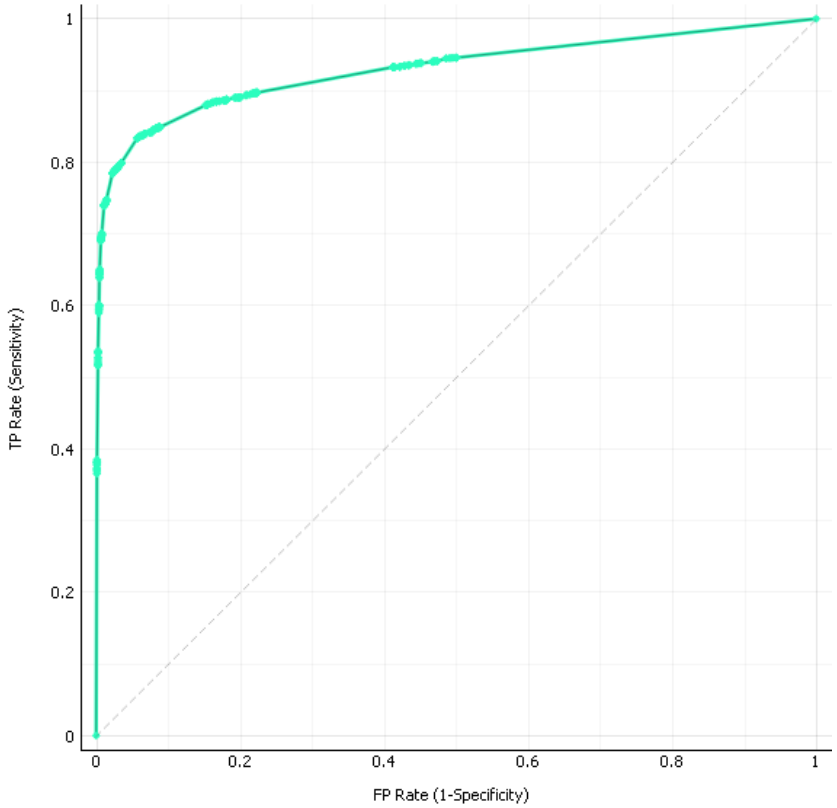


Evaluation

Time series cross-validation



ROC AUC on test set



Thank you!

dmitry.kolesnykov@gmail.com