Intelie by Viasat
Data Science Challenge

Project goal

Extract insights to help in the development of **Slip to Slip connection time** KPI.

Workflow

Exploratory data analysis

- Analyse time series-related properties in the features

Pre-processing

- Prepare data to modeling

Machine learning

- Identify when slip is **on** or **off**

Conclusion

- Analyse the pros and cons

Exploratory Data Analysis

Features description

Feature	Description	Stationary	Granger test	Outliers
BDEP	Bit depth	Yes	All features	0.7 %
TPO	Fluid flow	No	HL, WOB	3.0 %
HL	Hook load	No	BHT	0.3 %
BHT	Block position	No	HL, WOB	0.5 %
WOB	Weight on bit	No	BHT	0.3 %

Note: RPM, TOR and DEPT were removed due to null/constant value.

Story behind the data

Trip out operation

- Constant values for **RPM**, **TOR** and **DEPT** features.
- **HL** and **WOB** have an antagonistic behaviour.
- **BHT** quickly decreases when **on_slips**
- and slowly increases when **off_slips.**
- **BDEP** has a downward trend.
- **TPO** correlates with a change in seasonality.

Seasonality

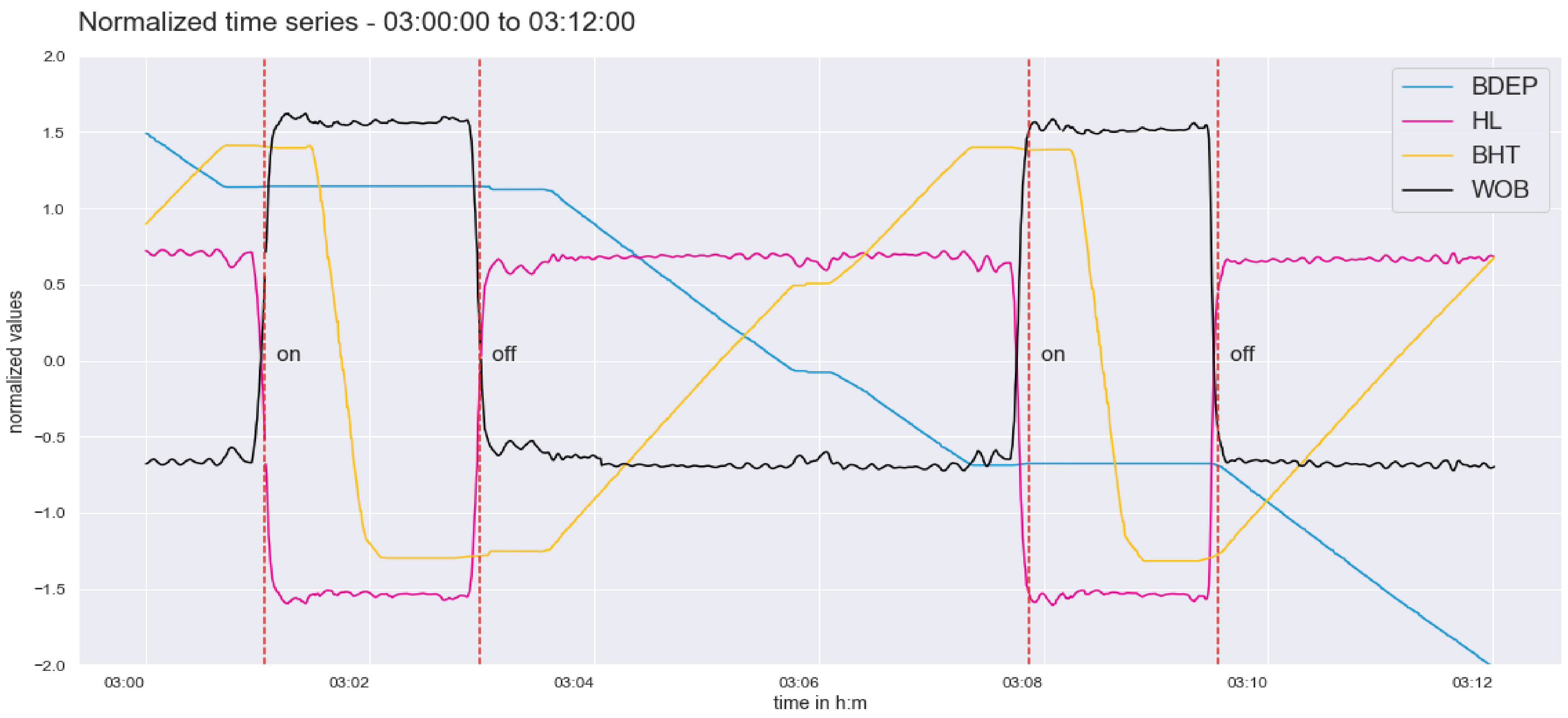
Average window time:

- on_slips = 2 minutes
- off_slips = 4 minutes

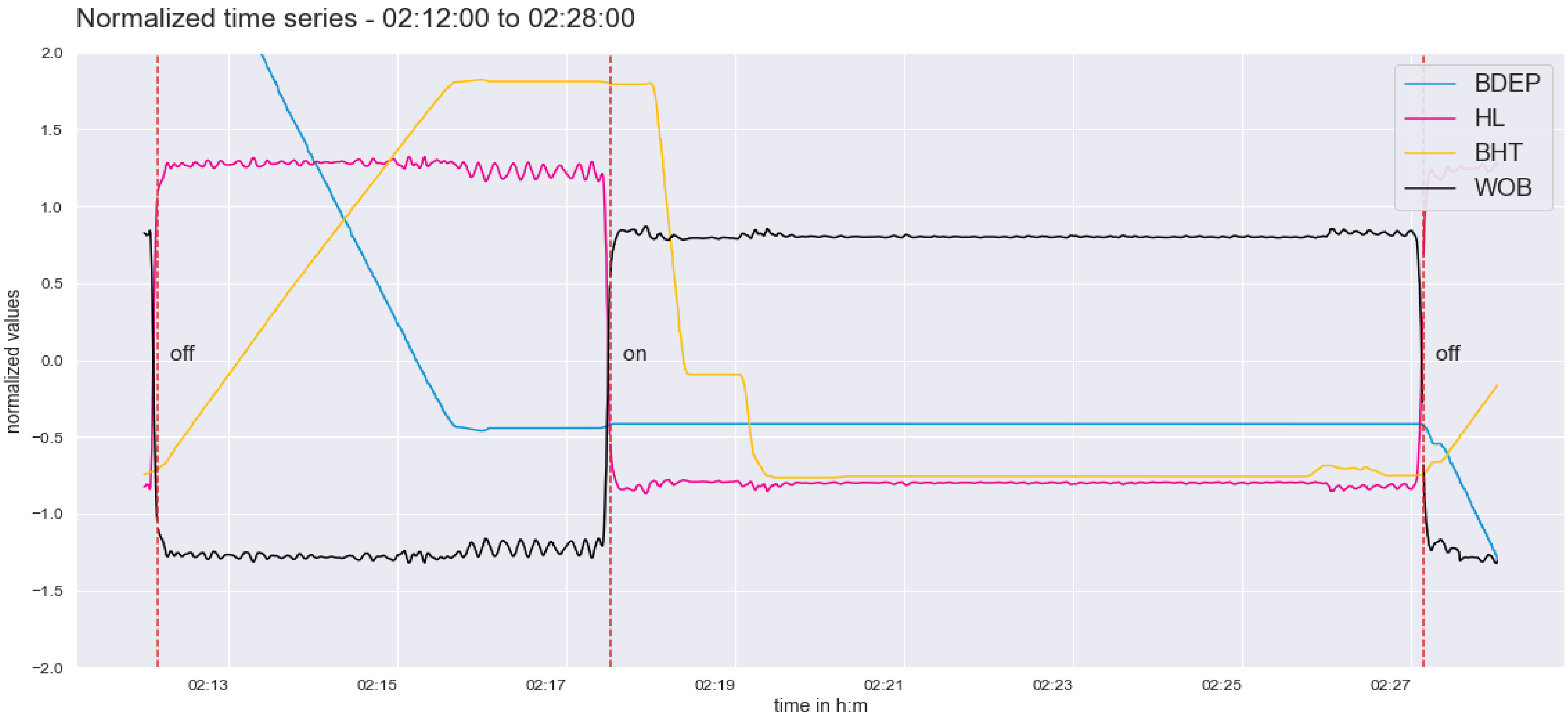
Outlier pattern from 02:12 to 02:28

- Precedes peak values of **TPO**

Time series pattern



Time series pattern



Preprocessing

Missing data

- Removed due to low percentage, 0.27%

Signal noise

- Smoothed with moving average

Empty values for Annotation

- Filled forward and then backwards

Preprocessing

Feature engineering

- Rolling mean, std and lag features

Split values for time series

- K-fold method for cross validation

Normalize values

- Standard scaling

Modeling

Random Forest

XGBoost

LSTM neural network

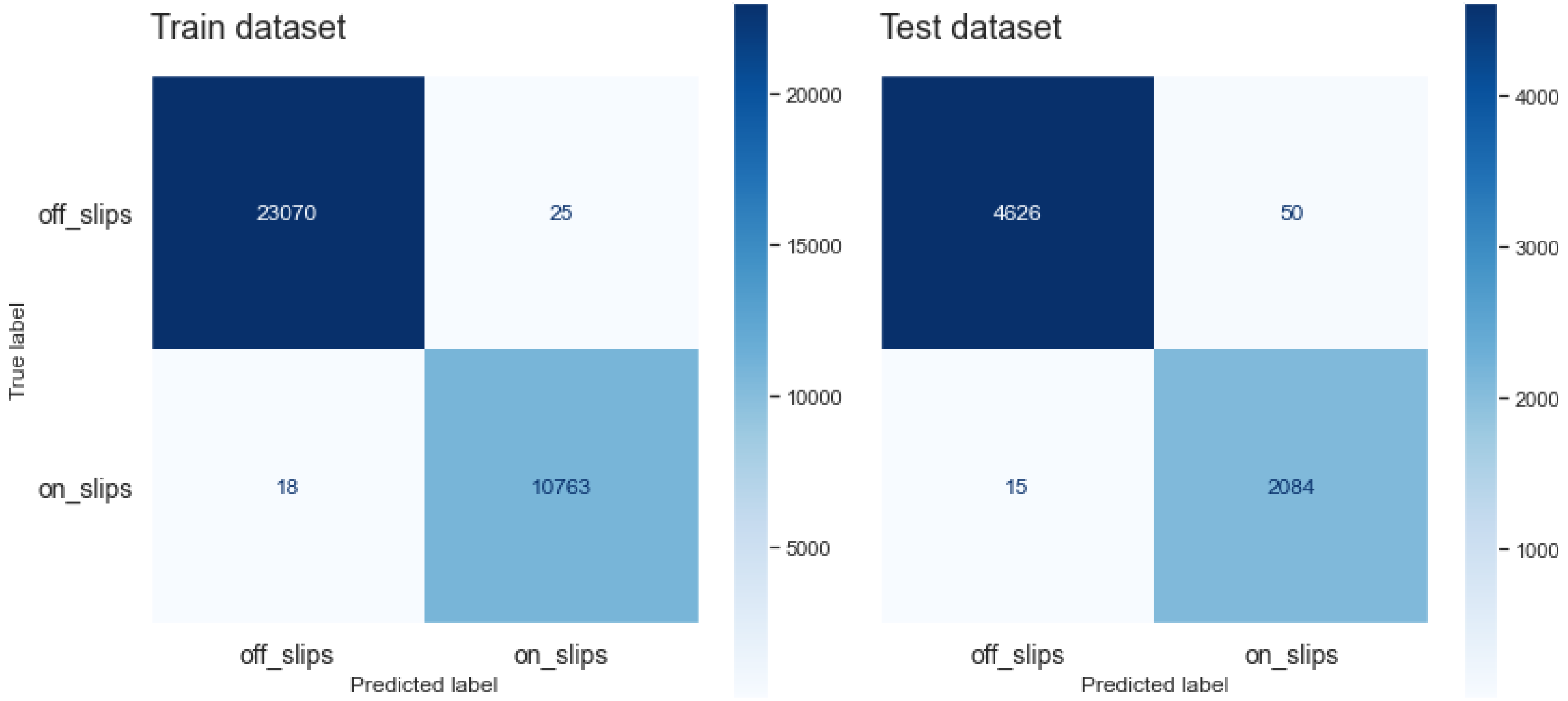
Baseline model

Random Forest Classifier with default params
to use as a benchmark.

Cross validation score:	0.9907
F1-score for **Train** dataset:	1.0
F1-score for **Test** dataset:	0.99

XGBoost evaluation

Confusion matrix



Model overfitting

Solutions to avoid overfitting:

- Use cross-validation ✓
- Apply regularization
- Collect more data

Model overfitting

Using regularization on Random Forest
had no impact.

Adding noise to the **Test data** lowered the
F1-score to 0.87 for **off_slips** and 0.53 for **on_slips**.

XGBoost Classifier

Early stop at 62º epoch

High score on both datasets: 0.97

2 high importance features:

- WOB ≅ 0.8
- HL ≅ 0.16

XGBoost Classifier

Predicting future variables with similar datasets produces good predictions.

When adding noise, the model complete fails to predict **on_slips** labels.

LSTM model

This model can handle long-term dependencies

Ability to capture seasonality and trends

Additional regularization techniques

LSTM model

Performance slight less overfitted
if compared to XGBoost

Still unable to predict label **on_slips** with
noise dataset.

LSTM evaluation

Conclusions

Both *XGBoost* and *LSTM* were good at predicting labels for **time series with the same distribution**.

It had no impact if the time between on/off slips changed. Just if heavy noises were introduced, or different patterns occurs.

Next steps

Try more robust regularization methods, or even increase layers for *LSTM*.

Collect more data in order to understand different patterns.

Thank you!

João Francisco Baiochi

Github: @Baiochi

E-mail: joao.Baiochi@outlook.com.br

Code in Jupyter notebook