VIASAT Data Science Technical Challenge

Raphael Crespo

Data Features

Features

index: timestamp in ms

* BDEP: bit depth in m

* TPO: fluid flow in gpm

* HL: hook load in klbf

* BHT: block position in m

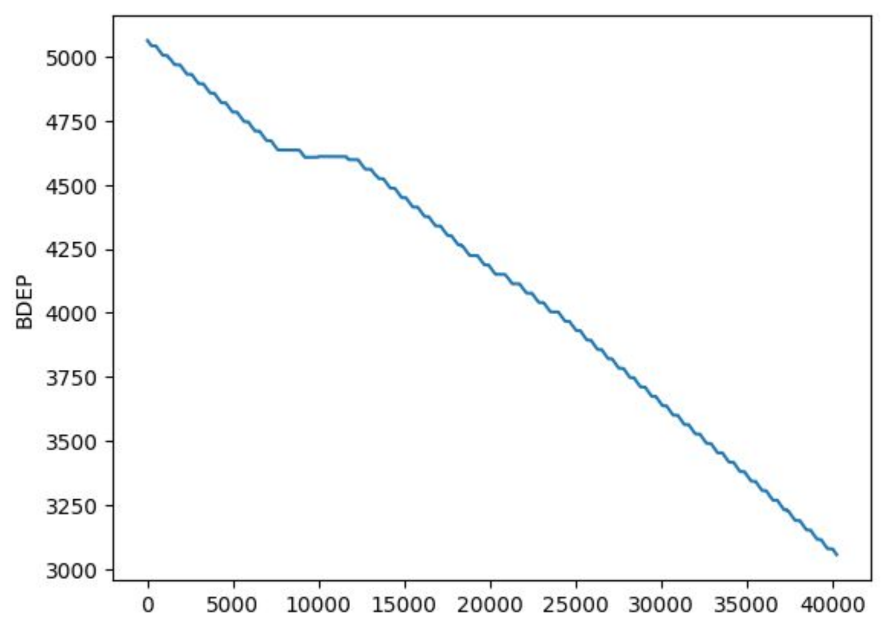
* RPM: rotary speed in rpm

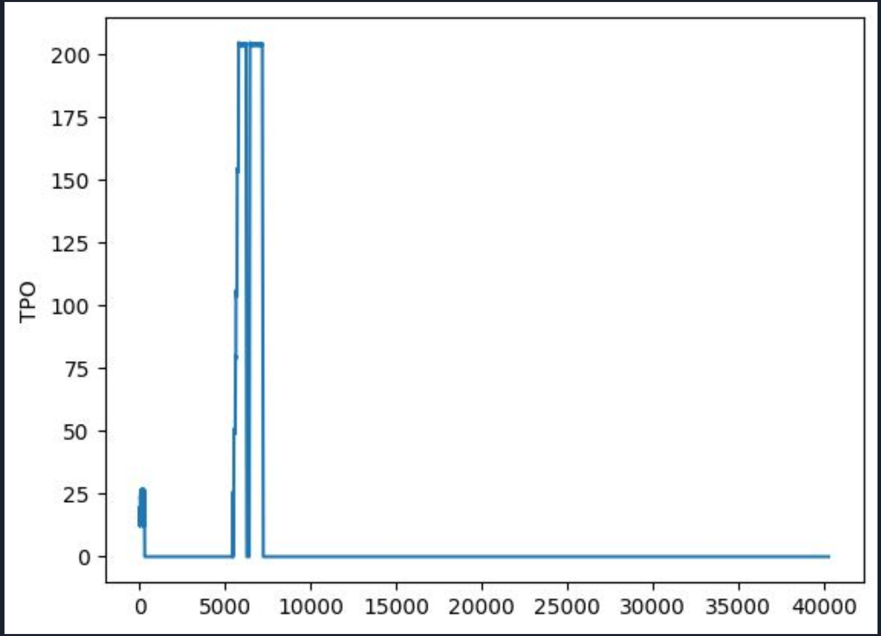
* TOR: torque in klbf-ft

* DEPT: hole depth in m

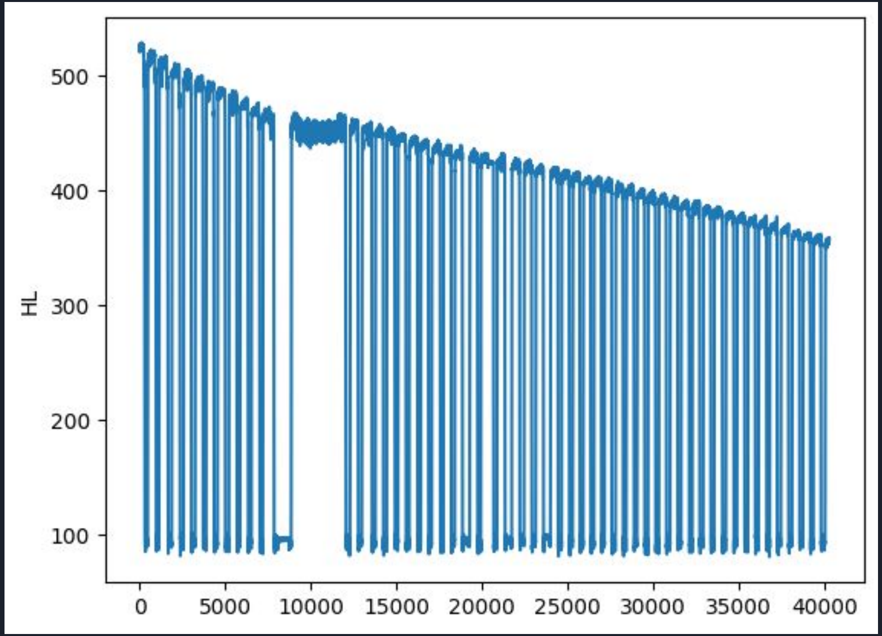
* WOB: weight on bit in klbf

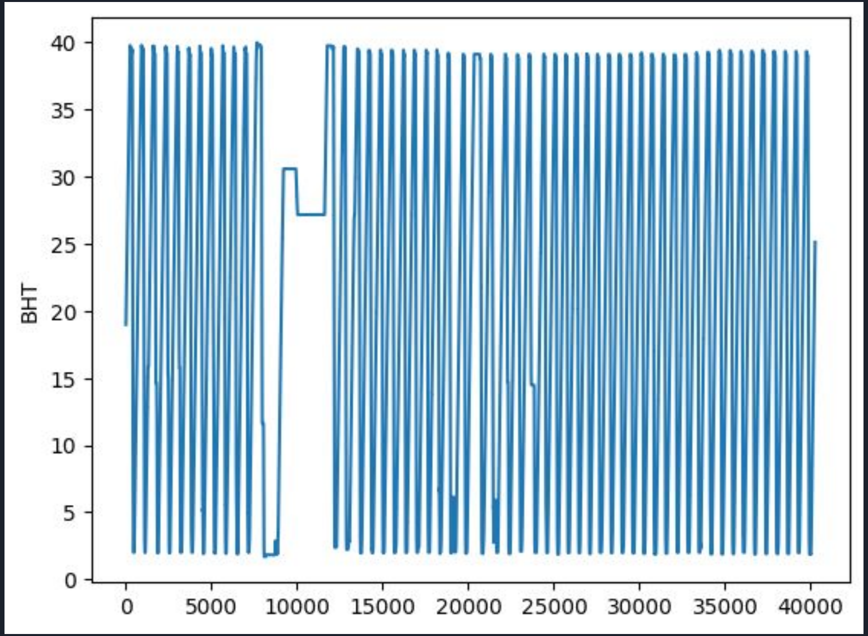
BDEP and TPO


BDEP
5000
4750
4500
4250
4000
3750
3500
3250
3000
0
5000
10000
15000
20000
25000
30000
35000
40000

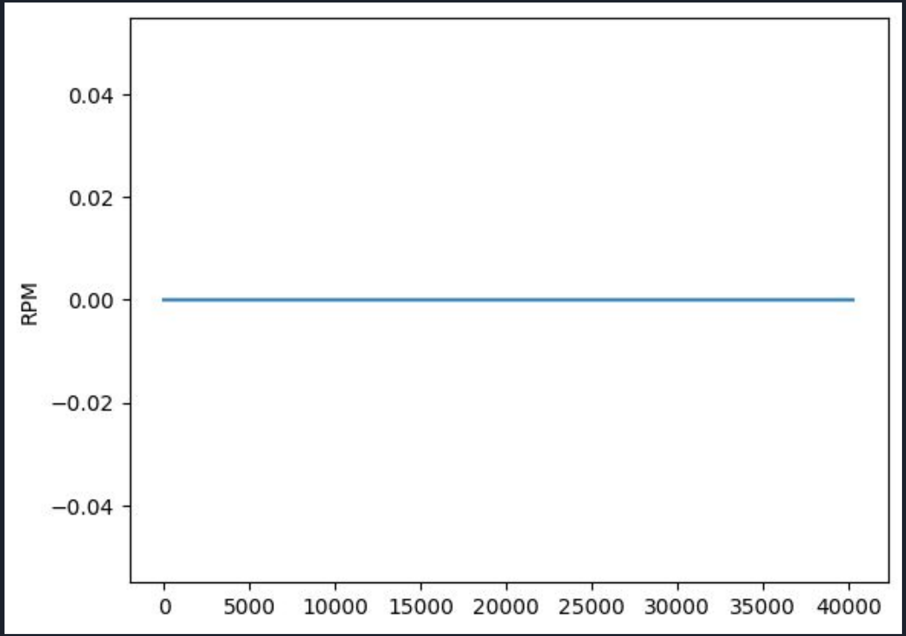

TPO
200
175
150
125
100
75
50
25
0
0
5000
10000
15000
20000
25000
30000
35000
40000

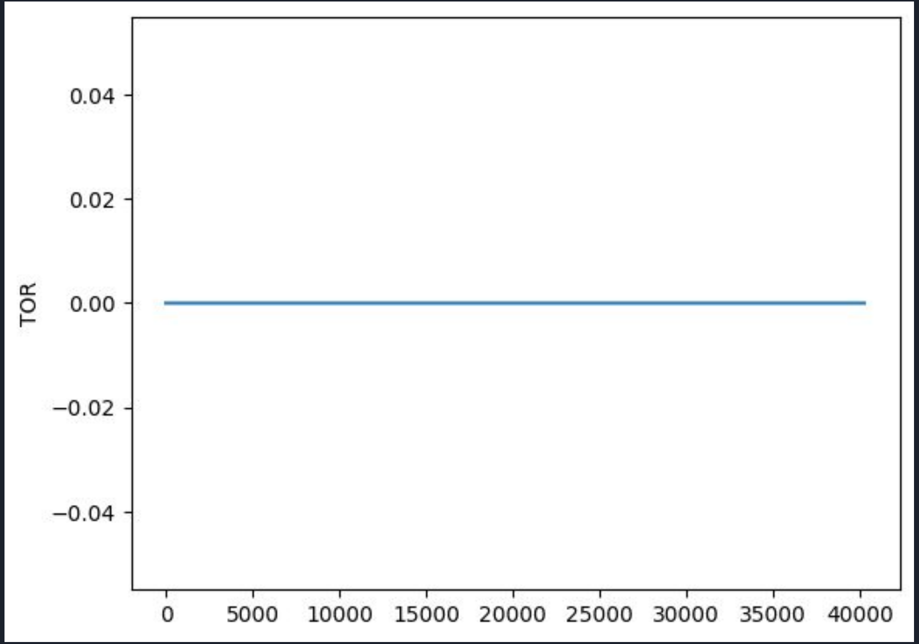
HL and BHT


HL
500
400
300
200
100
0
5000
10000
15000
20000
25000
30000
35000
40000

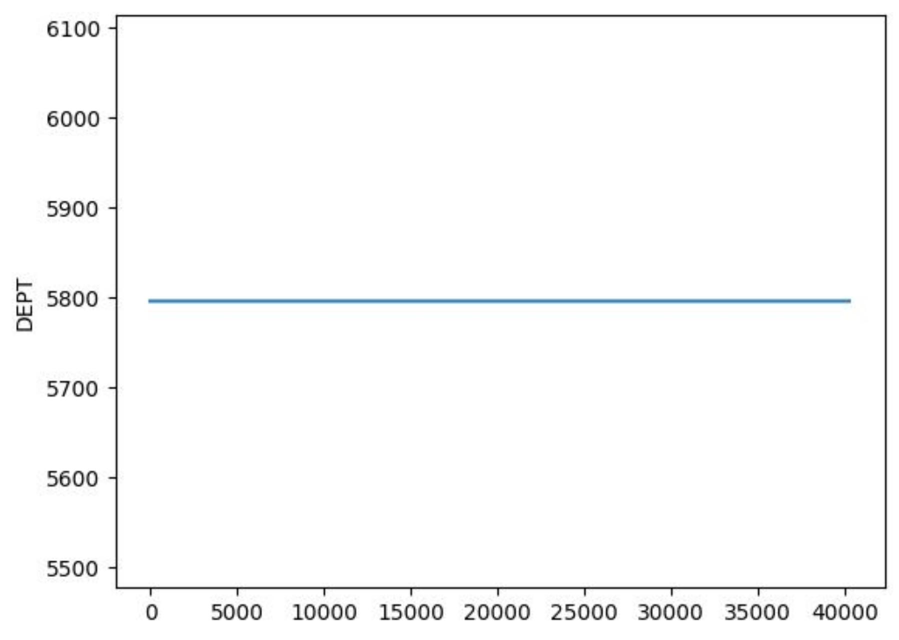

BHT
40
35
30
25
20
15
10
5
0
0
5000
10000
15000
20000
25000
30000
35000
40000

RPM and TOR

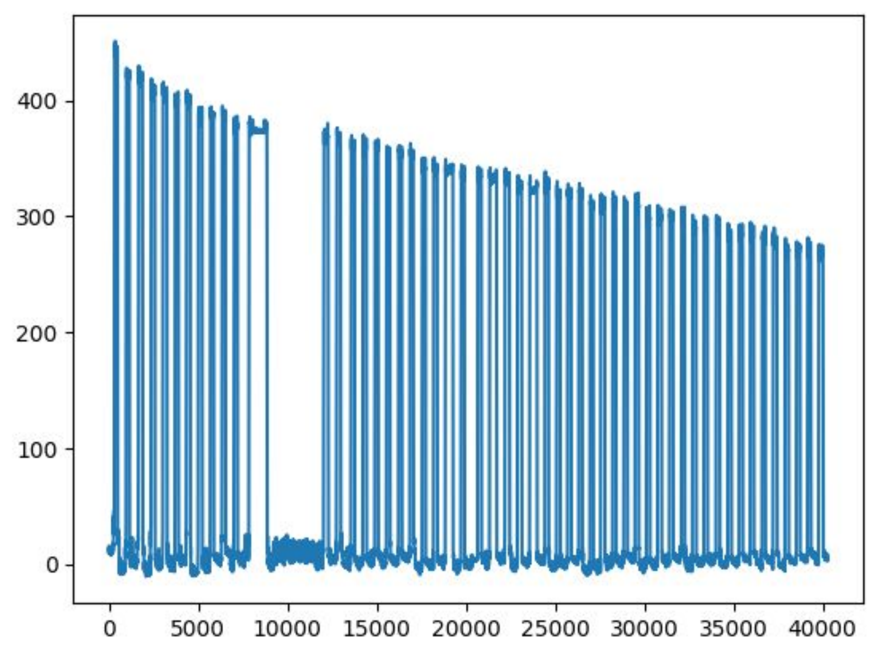

RPM
0.04
0.02
0.00
−0.02
−0.04
0
5000
10000
15000
20000
25000
30000
35000
40000


TOR
0.04
0.02
0.00
−0.02
−0.04
0
5000
10000
15000
20000
25000
30000
35000
40000

DEPT and WOB



DEPT
6100
6000
5900
5800
5700
5600
5500
0
5000
10000
15000
20000
25000
30000
35000
40000



400
300
200
100
0
0
5000
10000
15000
20000
25000
30000
35000
40000

Data Target

Target

Annotation: annotations

Binary variable

‘on_slips’

‘off_slips’

Data Preprocess

Handling Missing and Null Variables:

Target:

The chosen approach to address NaN values in the target variable ("Annotation") is to fill them with "on_slips" based on the information provided in the metadata file. According to the metadata, the value "off_slips" remains constant until the occurrence of the next annotation. Therefore, "on_slips" is used as a substitute for the missing values, ensuring consistency with the described behavior of the variable.

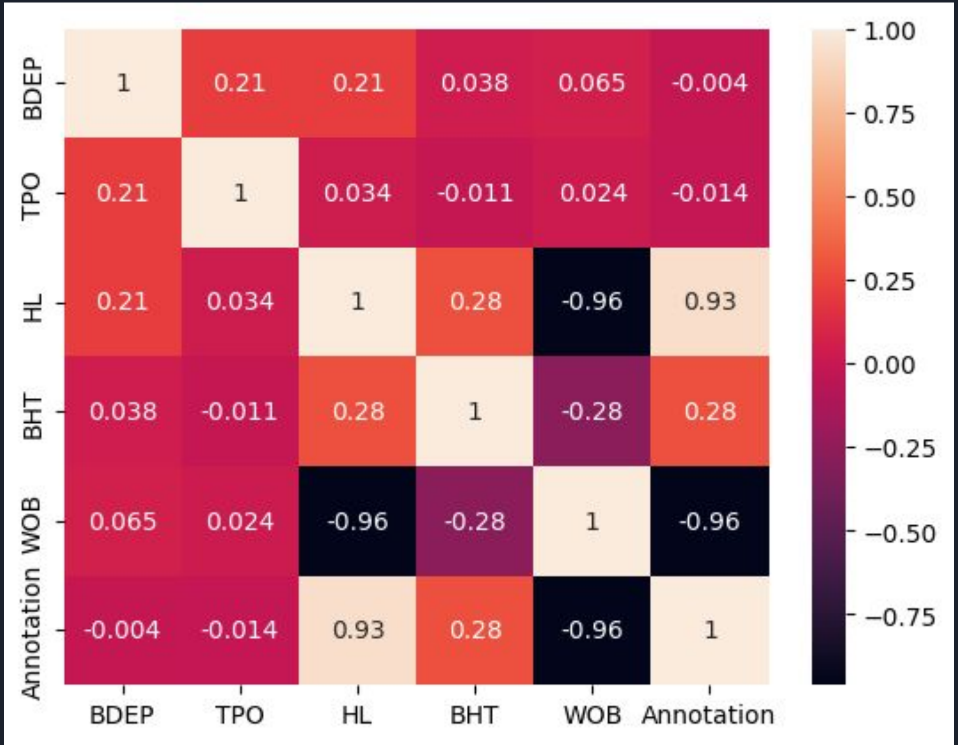
Features:

Applying the same logic as used for the target variable, the strategy employed to maintain time continuity of the variables involves using the "ffill" method to fill the NaN values. This approach ensures that missing values are filled with the last observed non-null value, thereby preserving the temporal sequence of the data. By forward-filling, each variable retains its most recent valid value, facilitating the analysis and interpretation of the dataset while preserving the integrity of the time series.

Handling Duplicated rows:

To deal with the duplicated rows it was kept only the latest records.

EDA

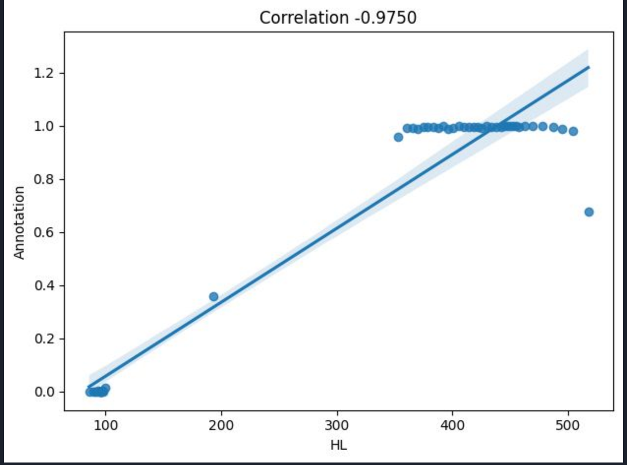


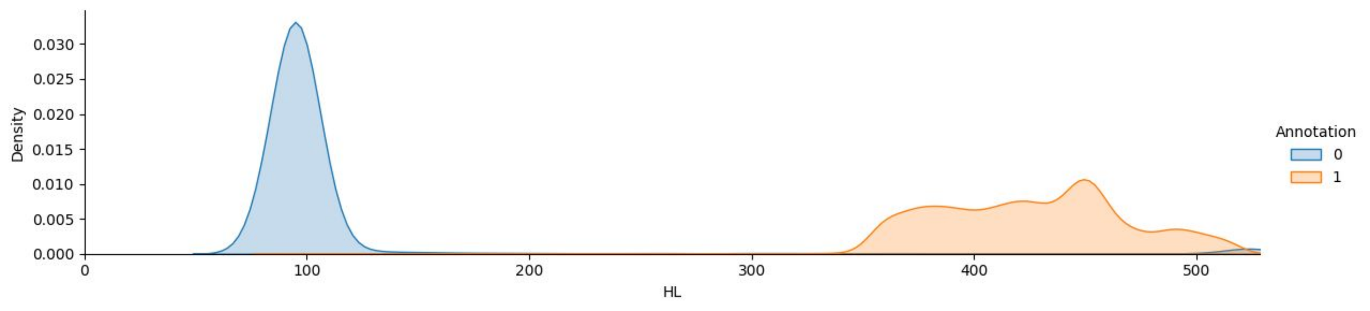
	BDEP	TPO	HL	BHT	WOB	Annotation
BDEP	1	0.21	0.21	0.038	0.065	-0.004
TPO	0.21	1	0.034	-0.011	0.024	-0.014
HL	0.21	0.034	1	0.28	-0.96	0.93
BHT	0.038	-0.011	0.28	1	-0.28	0.28
WOB	0.065	0.024	-0.96	-0.28	1	-0.96
Annotation	-0.004	-0.014	0.93	0.28	-0.96	1

EDA

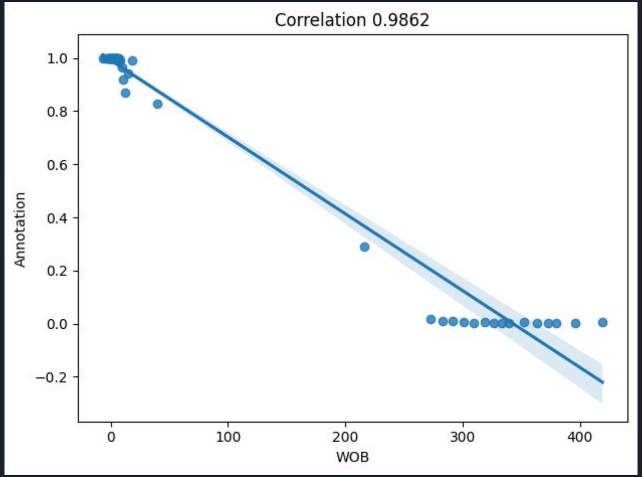
Based on the heat map analysis, it is evident that two features, namely WOB and HL, exhibit a strong correlation with the target variable, as well as with each other, with correlation coefficients approaching 1. On the other hand, the BDEP and TPO feature demonstrates a weak correlation value close to 0. Given this information, it is advisable to temporarily exclude the columns TPO and BDEP from further consideration.

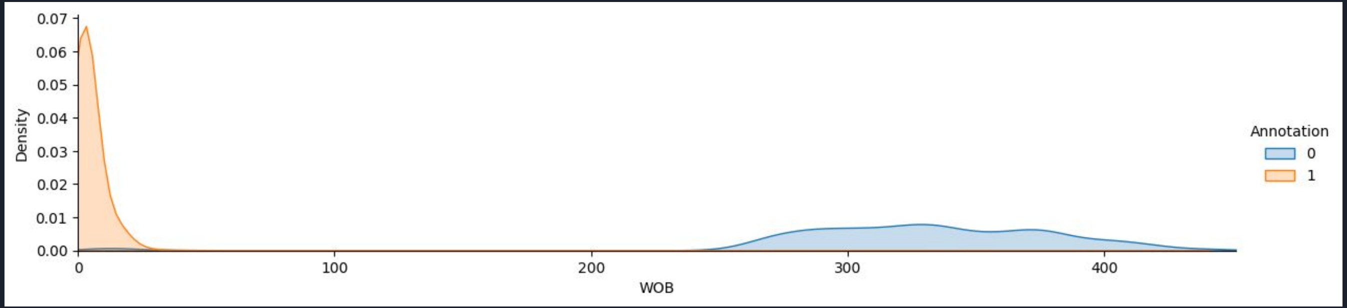
EDA - Overall Analysis of each Feature - HL



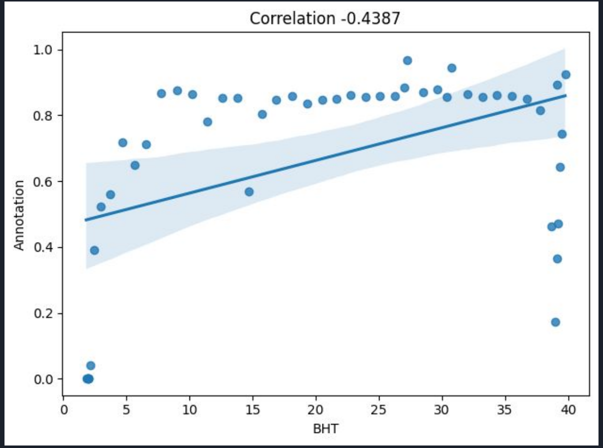


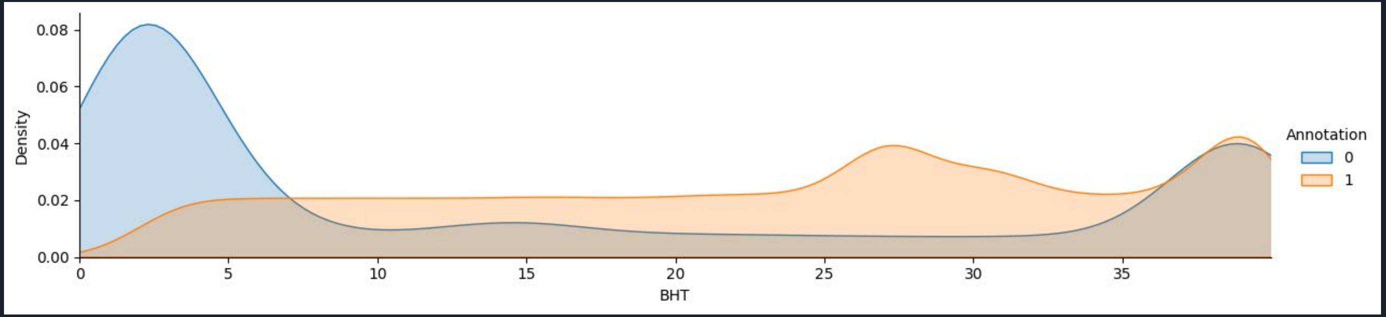
EDA - Overall Analysis of each Feature - WOB



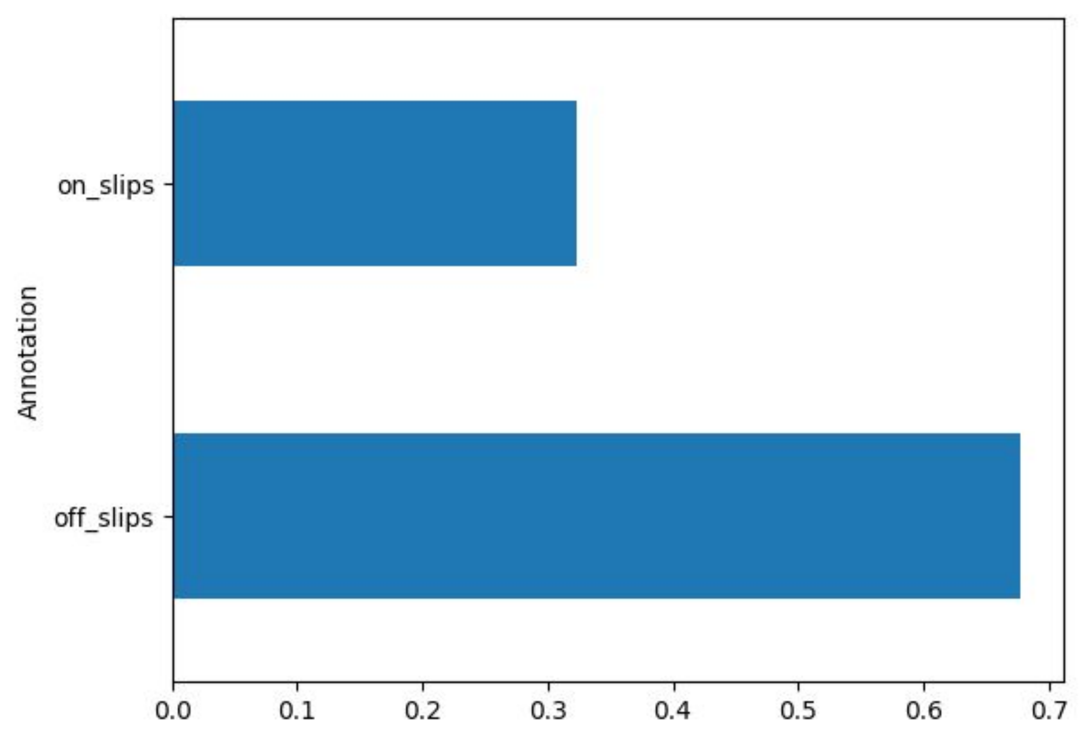


EDA - Overall Analysis of each Feature - BHT





EDA - TARGET



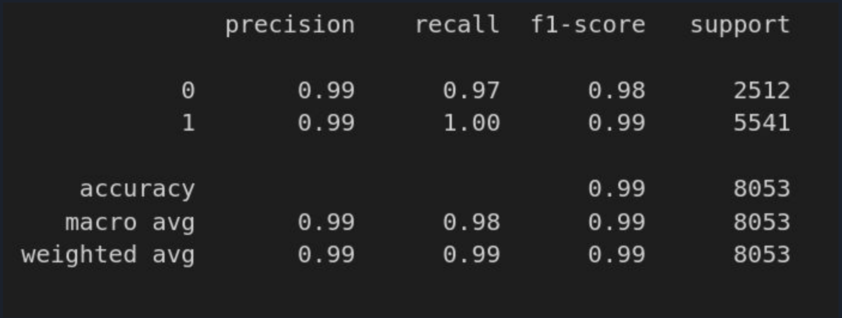
Modeling

Based on the findings from the exploratory data analysis (EDA), it was evident that an event occurred around instance 8000, causing a perturbation in the established data patterns across all variables. However, it appeared that the event subsided around instance 12000. To fully interpret the significance of this pattern within the data, the expertise of a specialist in this domain would be invaluable.

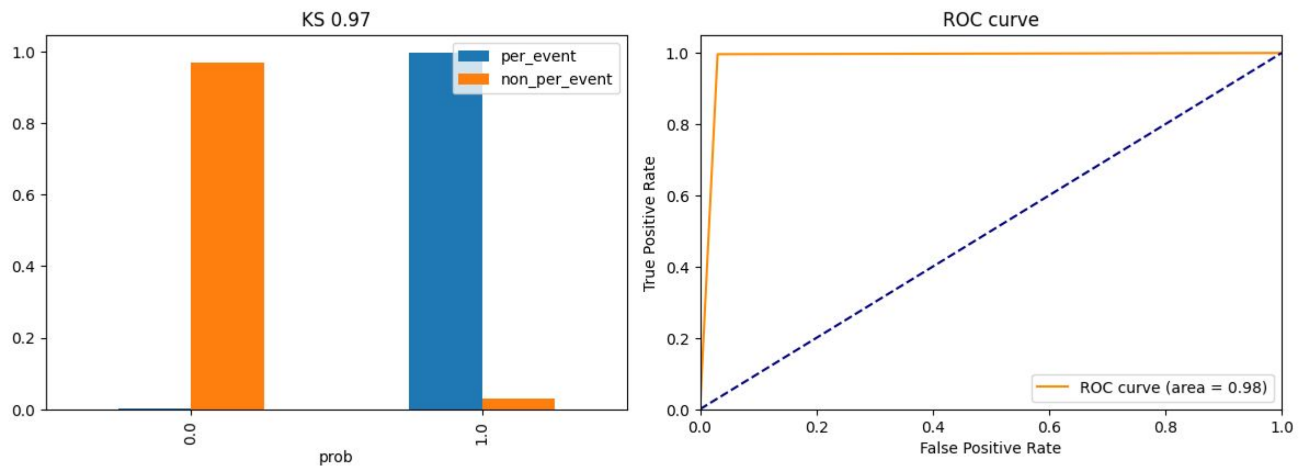
Considering this information, the modeling approach involved selecting features that exhibited a strong correlation with the target variable. Two different and straightforward machine learning models were then employed to gain initial insights into the behavior of the problem at hand.

Modeling

Logistic Regression

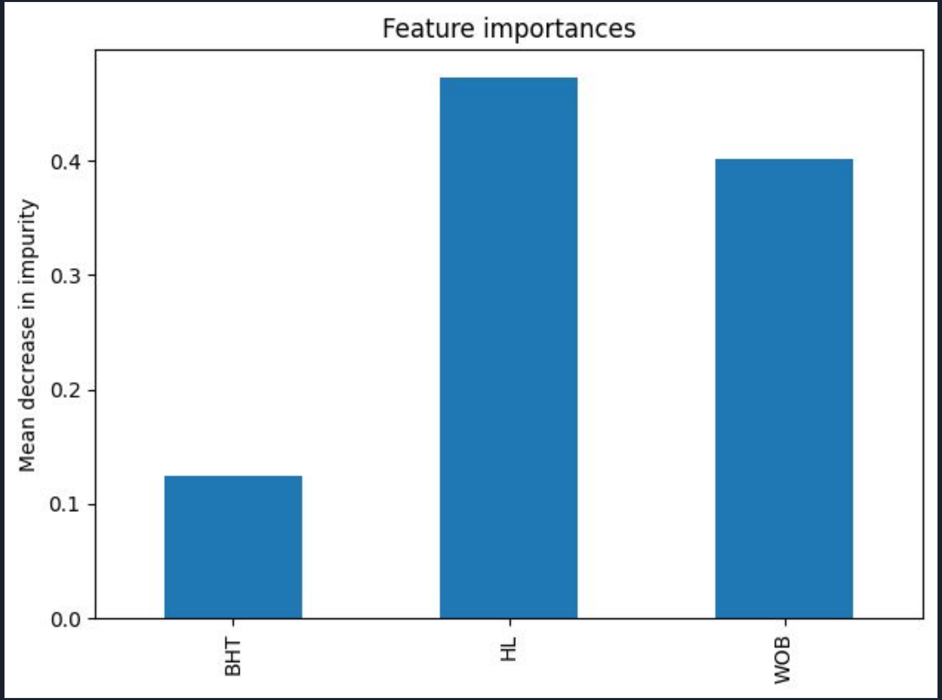


```
              precision    recall  f1-score   support

           0       0.99      0.97      0.98      2512
           1       0.99      1.00      0.99      5541

    accuracy                           0.99      8053
   macro avg       0.99      0.98      0.99      8053
weighted avg       0.99      0.99      0.99      8053
```



Modeling

Random Forest



Conclusions

The models exhibited impressive testing metrics, indicating the need for a closer and more in-depth assessment to understand the underlying reasons for the strong correlation between the target variable and the two features, namely HL and WOB. This assessment could potentially be best accomplished using a rule-based model, as it requires less computational power and offers a simpler explanation.

However, despite the advantages of a rule-based model, the final model selected and saved was the random forest. This decision was based on its superior performance metrics and its ability to provide interpretability through the feature importance method, which helps track the model's performance and stability.

To ensure the model's stability in a production environment, it is crucial to continuously monitor the model's metrics and the data's statistical characteristics. Metrics such as the Population Stability Index (PSI) and the Characteristics Stability Index can be used over time to assess and maintain the model's stability.