Predictive Maintenance

NASA Turbofan Engine Fault

Project Goals

Accurately predict “Remaining Useful Life” of the equipment.

The Problem…

Predict the number of operational cycles before a failure in the test set.

In other words, predict the number of operational cycles after the last cycle that the engine will continue to operate.

About The Data

- Each time series is from a different engine
- Each engine starts with different degrees of initial wear and manufacturing variation which is unknown to the user
- This wear and variation is considered normal
- There are three operational settings that have a substantial effect on engine performance
- The engine is operating normally at the start of each time series, and develops a fault at some point during the series.

ETL Strategy

We acquired the dataset from: https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#turbofan.

We used FD001 data, which consisted of 20630 rows and 28 columns that were made up of:

- Machine Unit Number
- Machine Cycle Count
- Operational Settings(3)
- Sensor measurements (20)

ETL Strategy Cont…

The data was stored in an S3 bucket, which we were able to access through a Notebook Instance in Amazon SageMaker. Once in the Notebook Instance the data was then prepared(cleaned/transformed) for analysis:

- After columns were renamed:
  - data_column = ["unit number","time","opt_setting1","opt_setting2","opt_setting3","t2","t24","t30","t50","p2","p15","p30","nf","nc","epr","ps30","phi","nrf","nrc","bpr","farb","htbleed","nf_dmd","pcnfr_dmd","w31","w32"]
- After columns were removed:
  - Data_column = ['unit number','time','opt_setting1','opt_setting2','opt_setting3','epr','t2','farb','p2','nf_dmd','pcnfr_dmd']

ETL Strategy Cont…

Then the provided RUL file dataset was added to the modified data frame and we began the EDA. The following images show these results.

- **In the training set, the fault grows in magnitude until system failure.**

Unit	Time	Feature 1	…	Feature N		RUL	
1	1	xxx	…	yyy		100	
1	2	xxx		yyy		99	
1	…	…	…	yyy	+	…	Inverse of cycle
1	100	xxx	…	yyy		1	
…							

ETL Strategy Cont…

However for the test data,

- **In the test set, the time series ends some time prior to system failure.**
- **RUL of the stop point is in a separate file (rul_xxx.txt)**

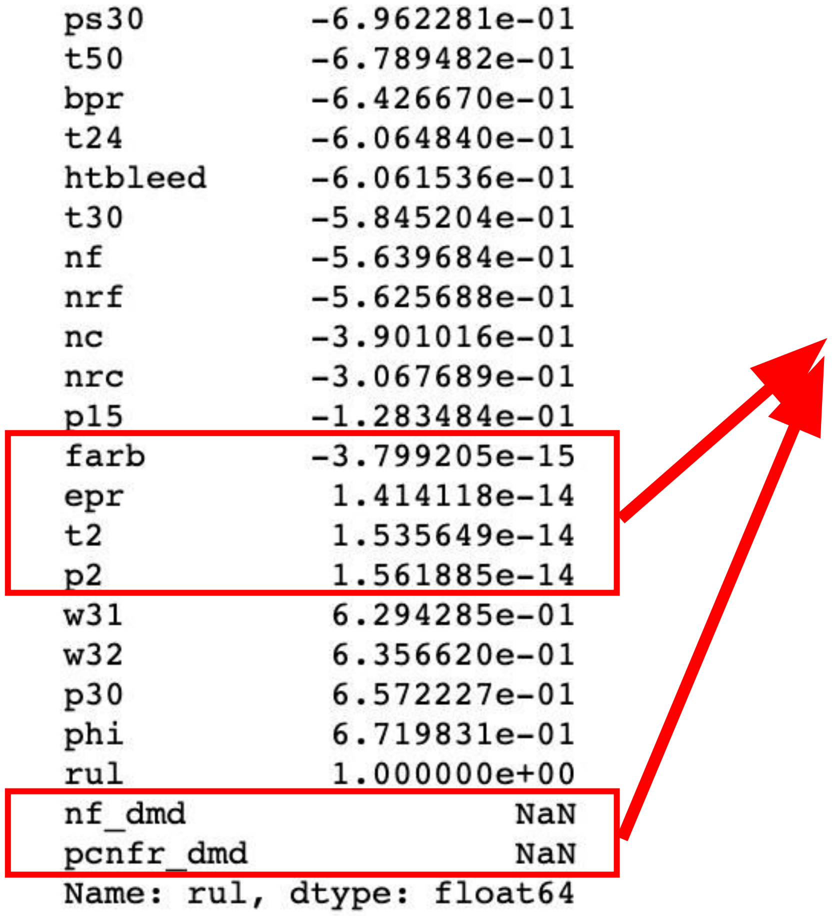
Actual RUL
from the file
-> 20

Unit	Time	Feature 1	…	Feature N
1	1	xxx	…	yyy
1	2	xxx		yyy
1	…	…	…	yyy
1	100	xxx	…	yyy
…				

+

RUL	
100 + 20 -1	119
99 + 20 - 1	118
…	…
1 + 20 -1	20

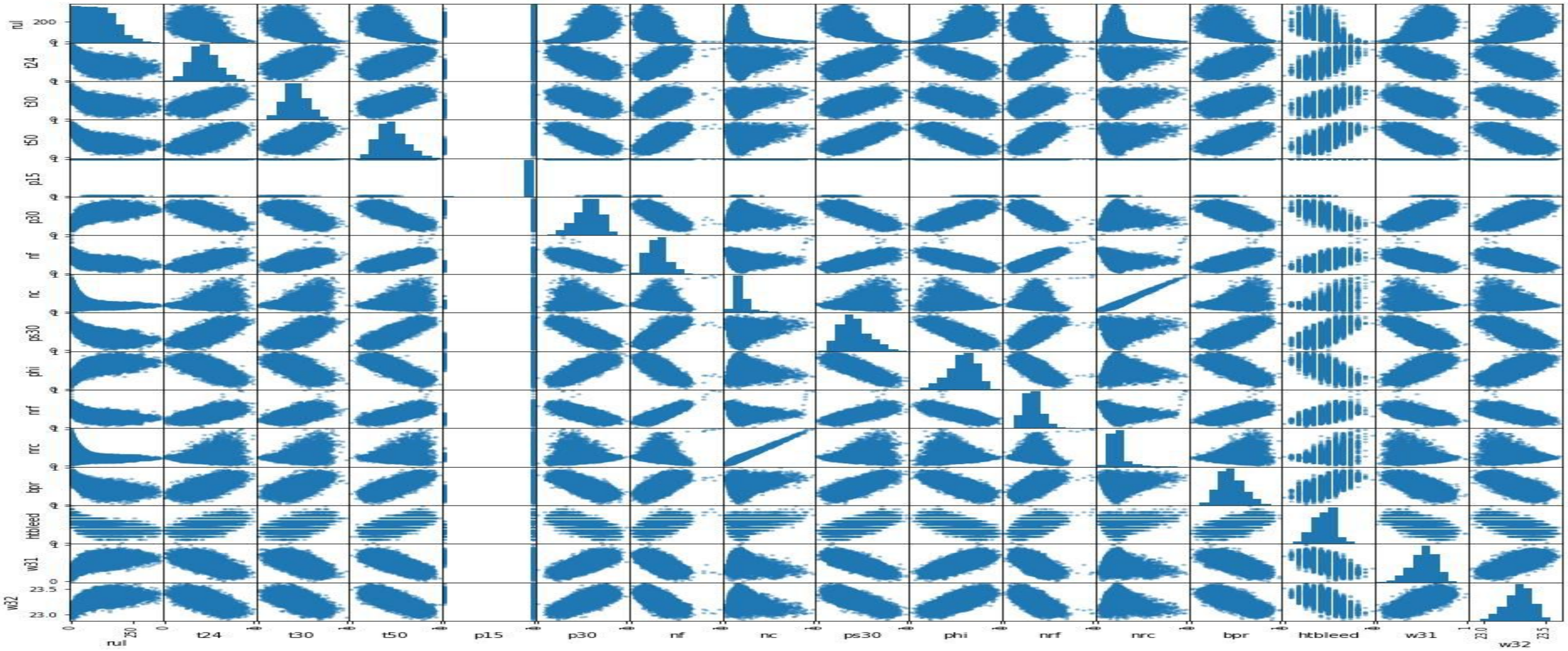
Correlation in respect to RUL



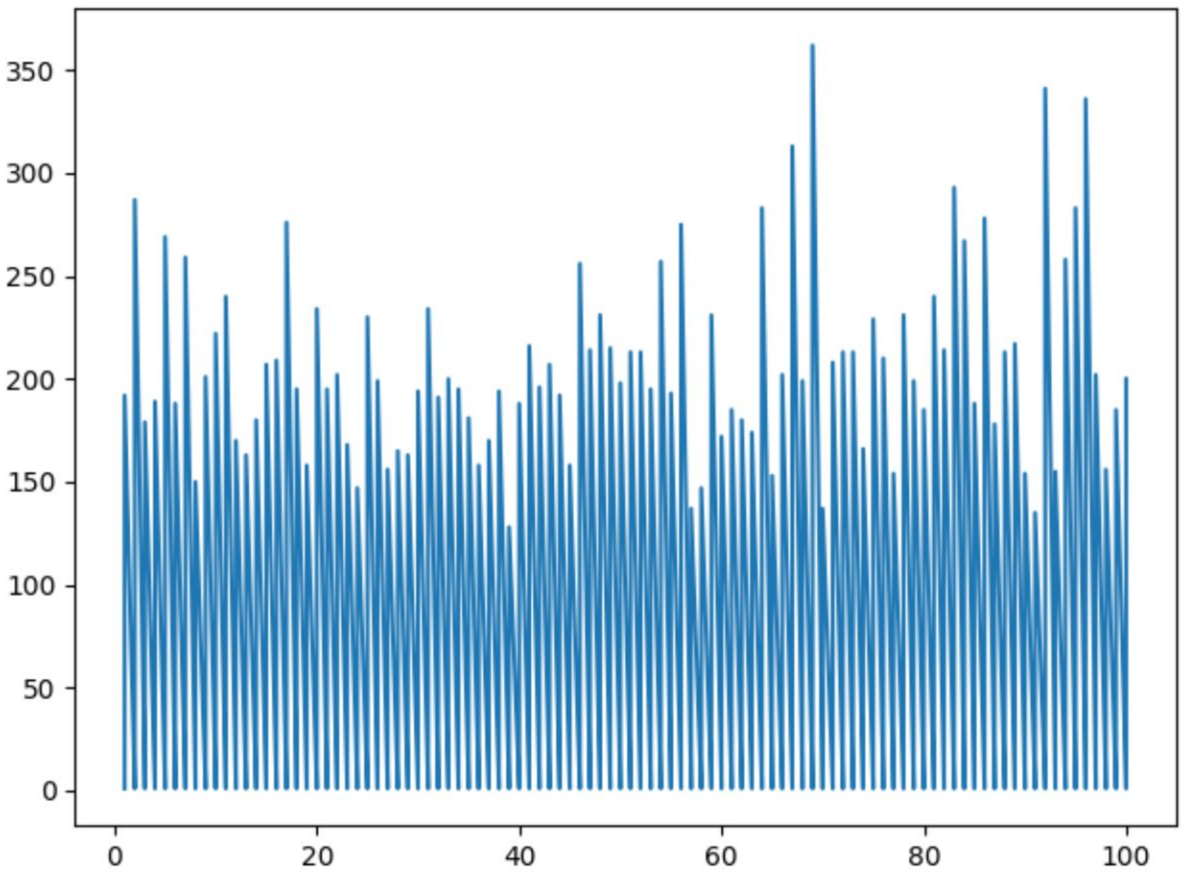
```
ps30          -6.962281e-01
t50           -6.789482e-01
bpr           -6.426670e-01
t24           -6.064840e-01
htbleed       -6.061536e-01
t30           -5.845204e-01
nf            -5.639684e-01
nrf           -5.625688e-01
nc            -3.901016e-01
nrc           -3.067689e-01
p15           -1.283484e-01
farb          -3.799205e-15
epr            1.414118e-14
t2             1.535649e-14
p2             1.561885e-14
w31            6.294285e-01
w32            6.356620e-01
p30            6.572227e-01
phi            6.719831e-01
rul            1.000000e+00
nf_dmd                  NaN
pcnfr_dmd               NaN
Name: rul, dtype: float64
```

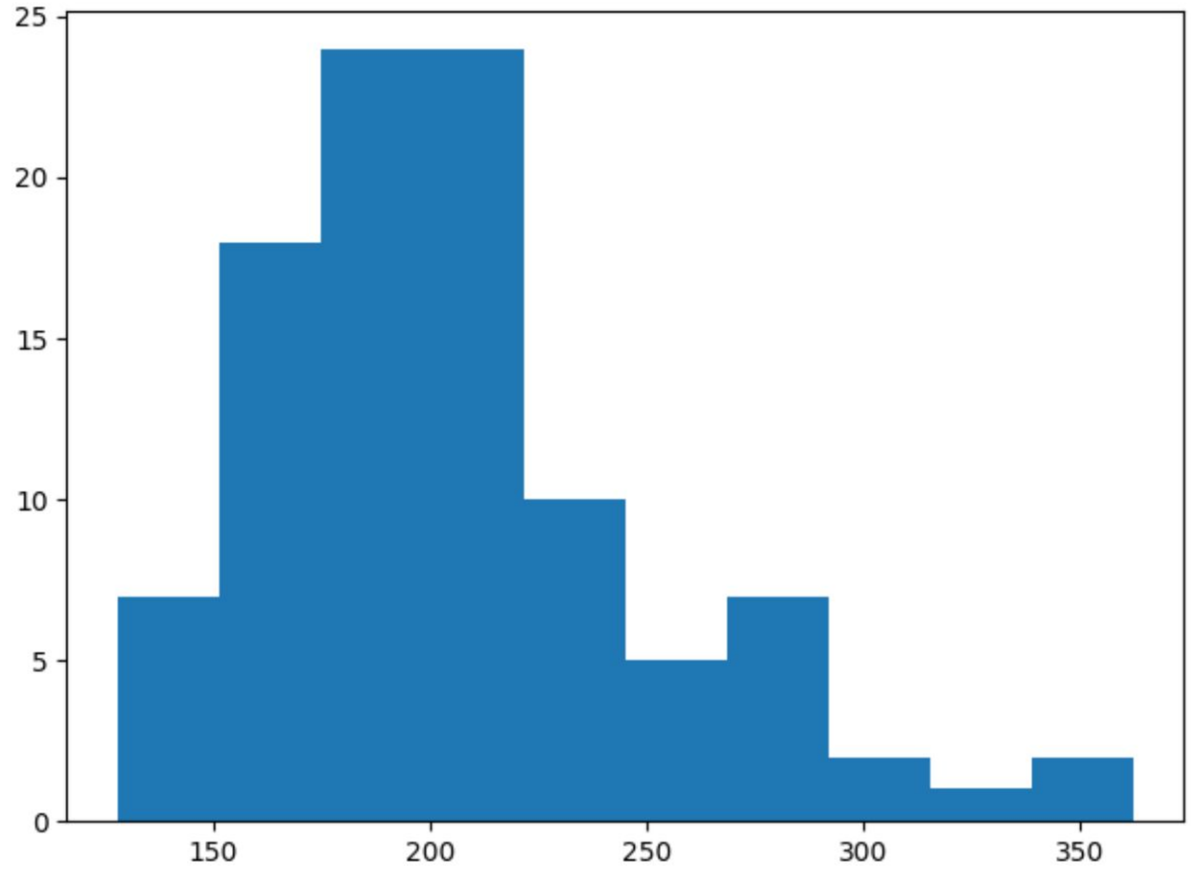
Low-sensitive measurement according to Pearson correlation RUL

Correlation via Matrix Scatterplot

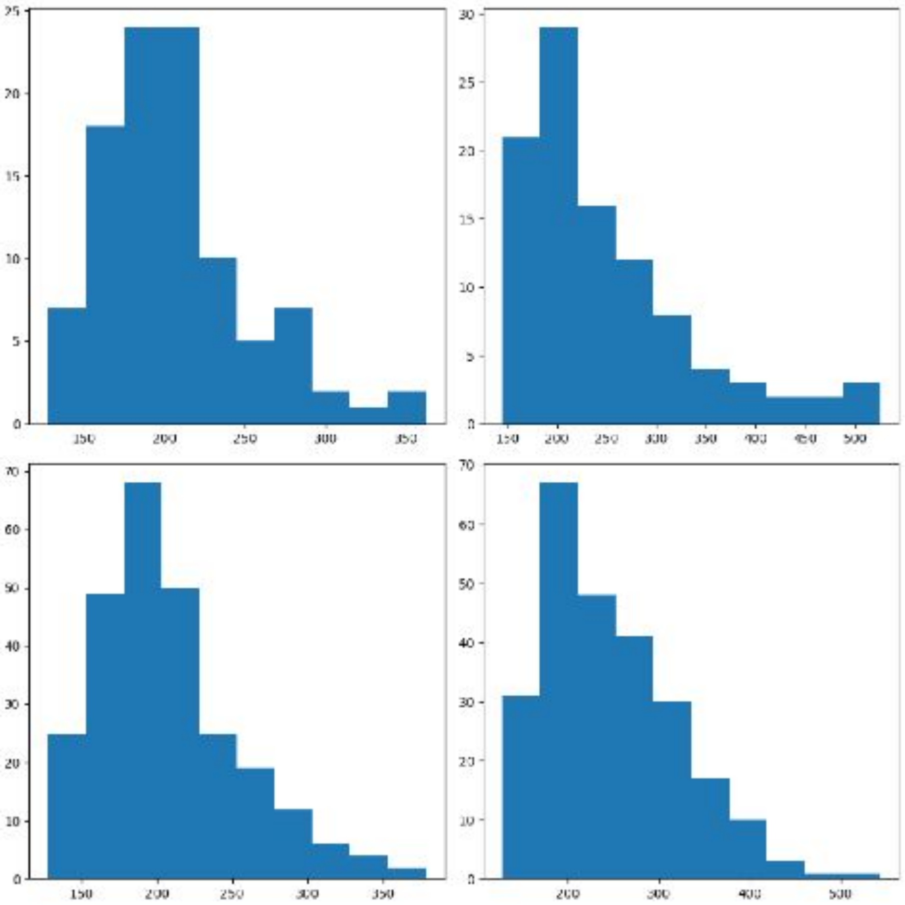


Remaining Useful Life Analysis



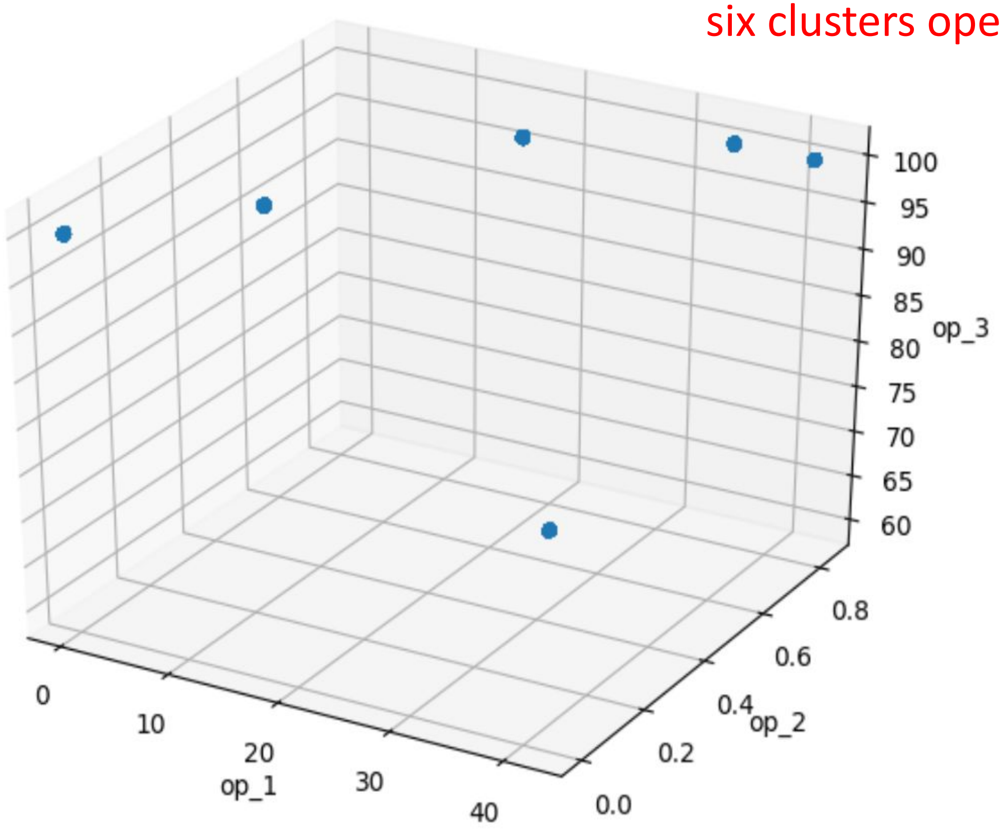


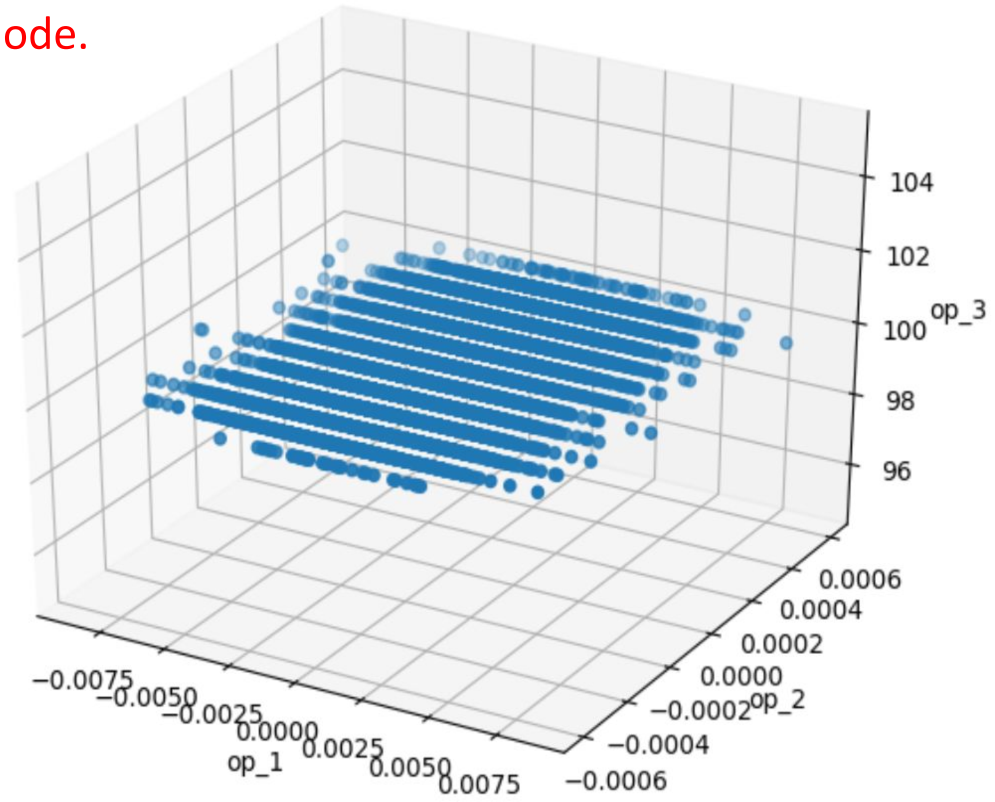
Remaining Useful Life Analysis cont…



Remaining Useful Life Analysis cont…

three operational settings, with
six clusters operations mode.

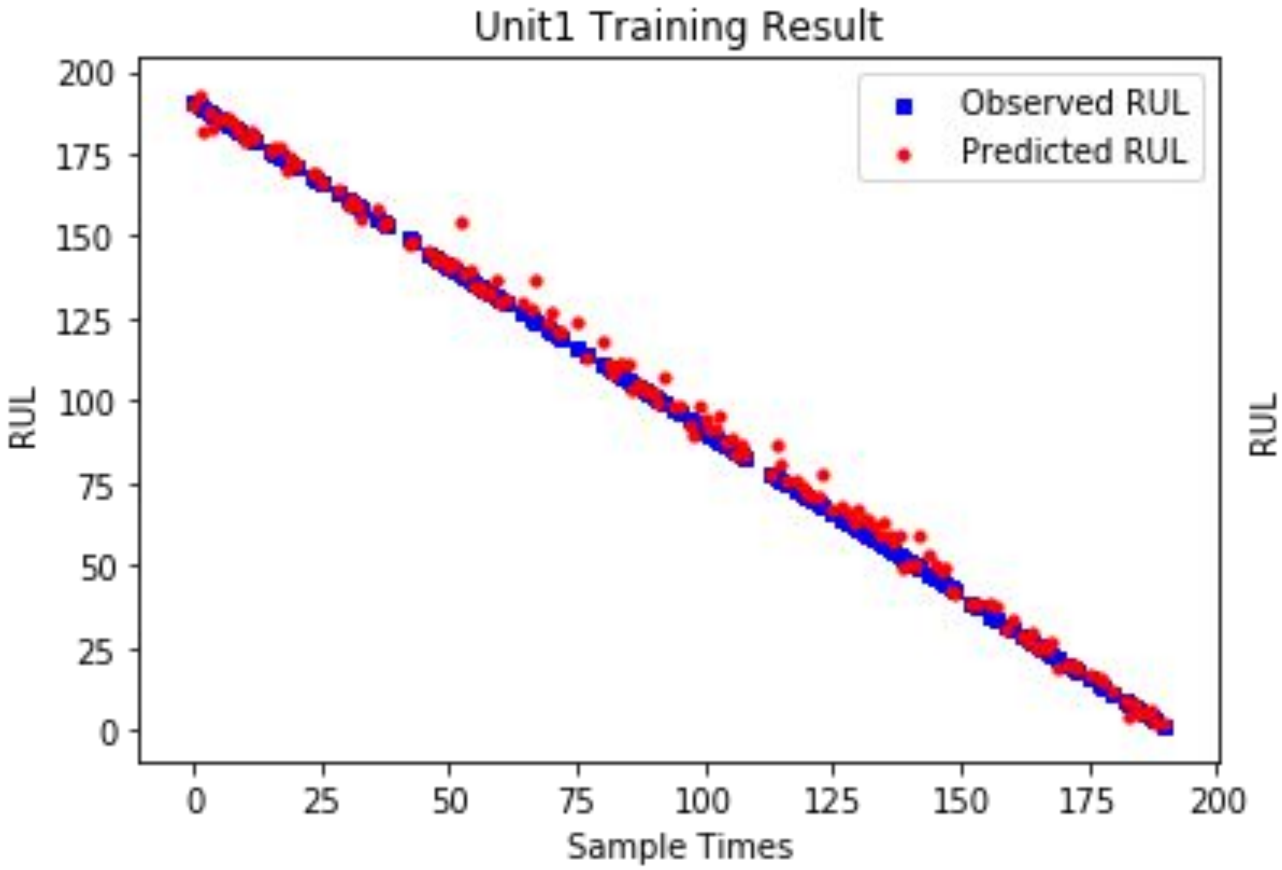


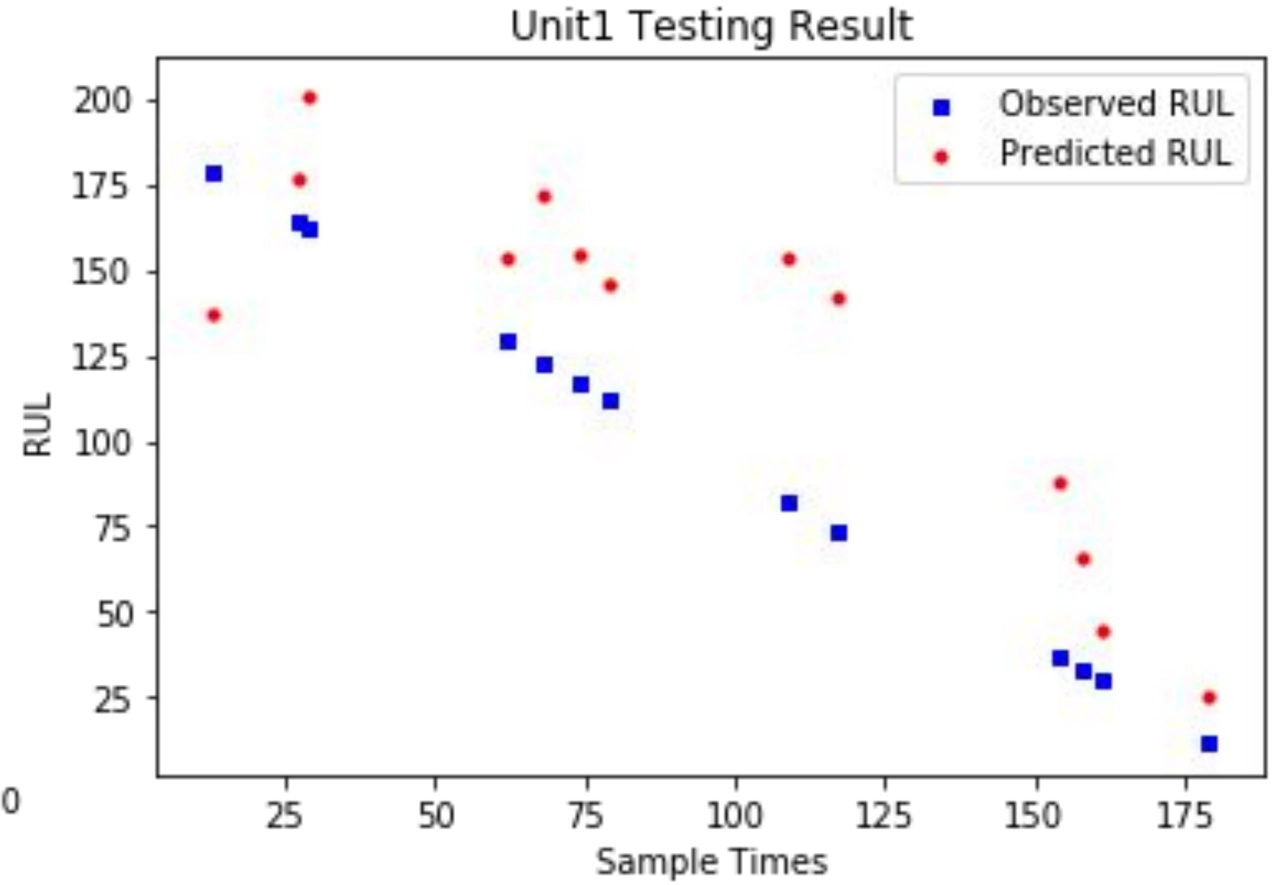


Model Strategy

We decided to utilize XGBOOST Linear Regression Algorithm to predict the ”Remaining Useful Life” of the machinery.

Results



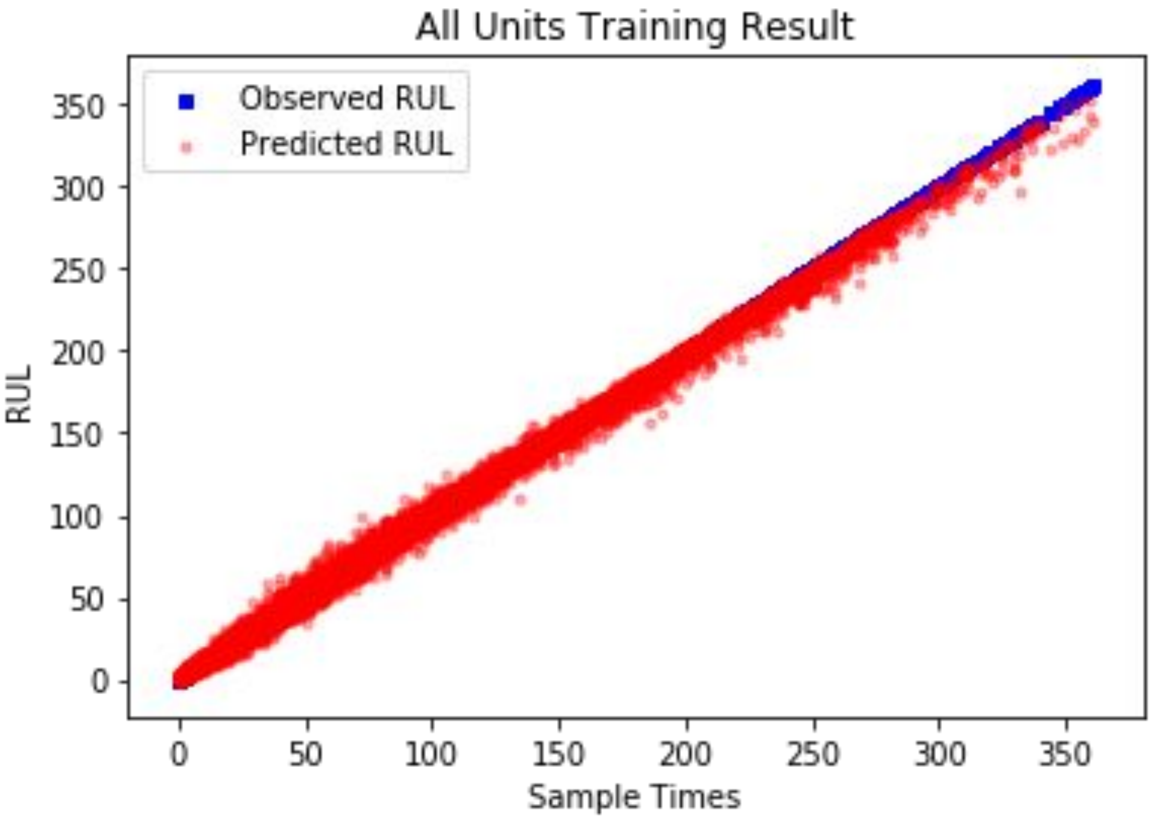


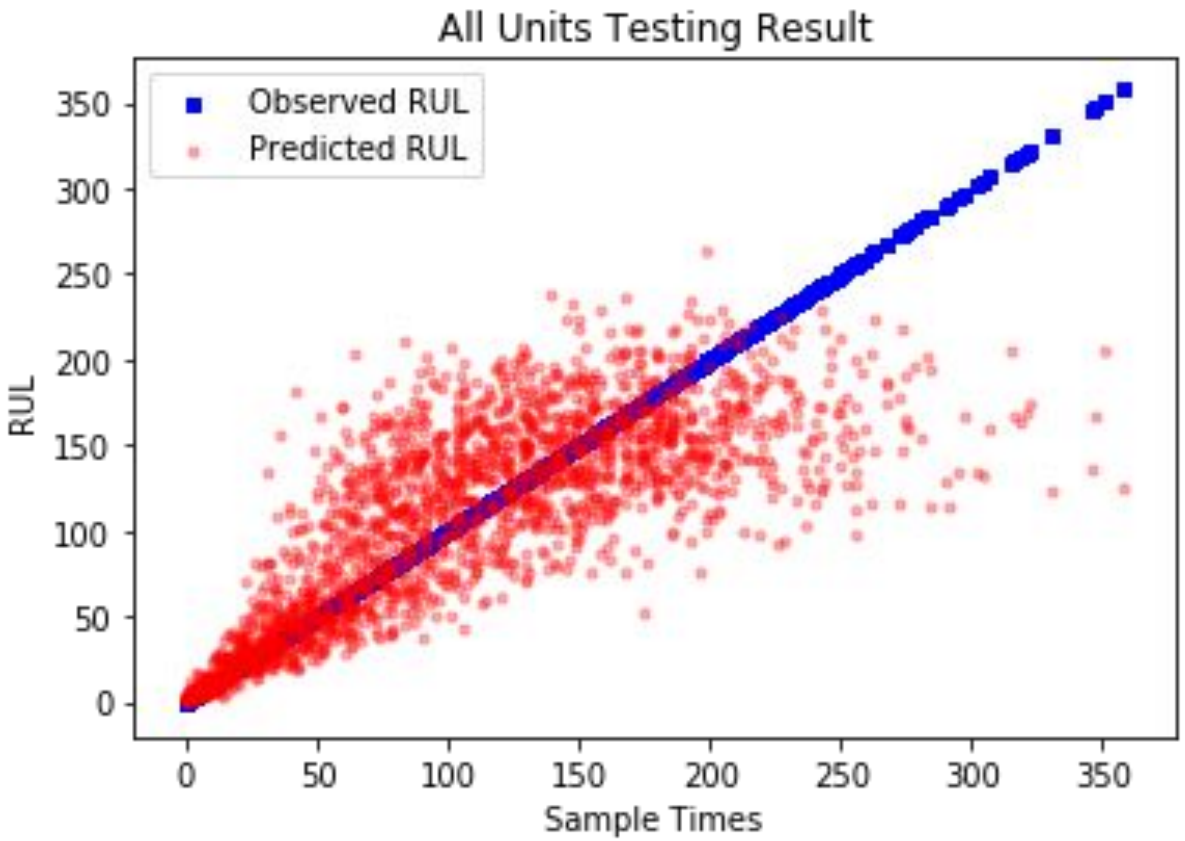
Training RMSE: 4.097

Testing RMSE: 42.57

The model is trained and predicted using data of 100 units in train_FD001.txt, only unit1 data was plotted

Results





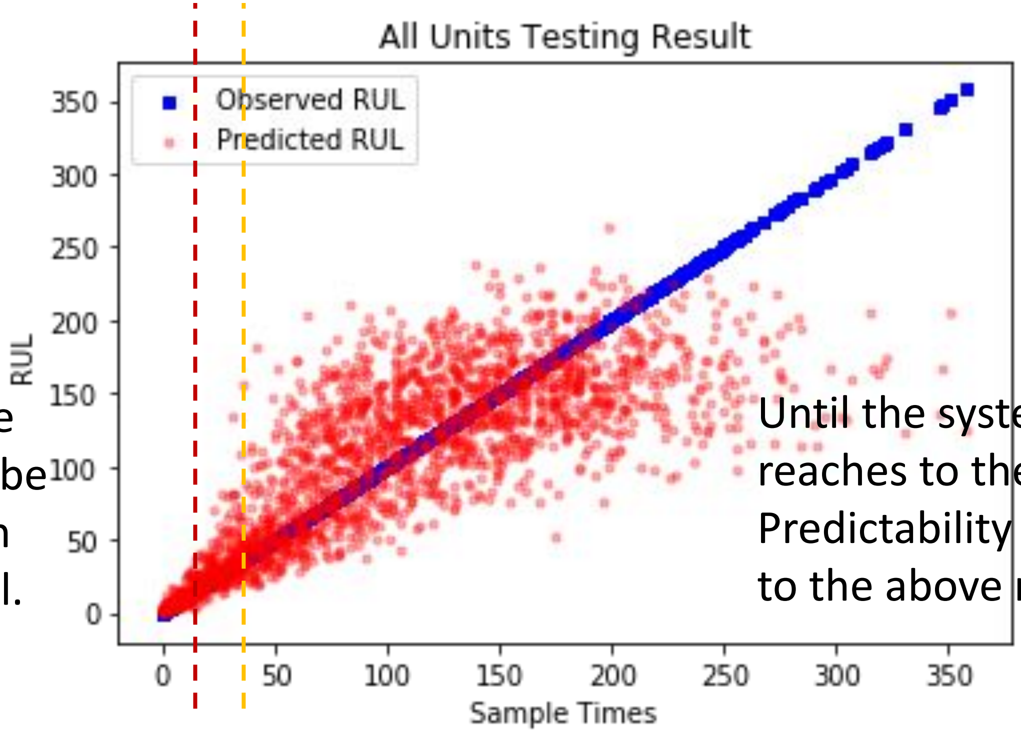
Training RMSE: 4.097

Testing RMSE: 42.57

It can be observed that when RUL reaching to breaking point, the prediction is closer to true RUL.

Conclusion

The engine is operating normally at the start of each time series, and develops a fault at some point during the series.



However, the remaining useful life is at certain point (e.g. 50), it can be seen that prediction saturate with quite accurate confidence interval.

Until the system degradation reaches to the certain point, Predictability is not quite high due to the above reason.

Action required

Lessons Learned

- The training dataset seems to lean towards creating a model that is overfit.
- Based on the nature the problem I assumes that a binary classification model would be best.
- Understanding the data is fundamental before you start the process.