SNIA DEVELOPER CONFERENCE
SDC 24
BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Storage Device Quality Control and Supply Chain Management Using Dual Machine Learning Models

Yongjin Choi, Mourad Larbi, Chris Cheng
Hewlett-Packard Enterprise

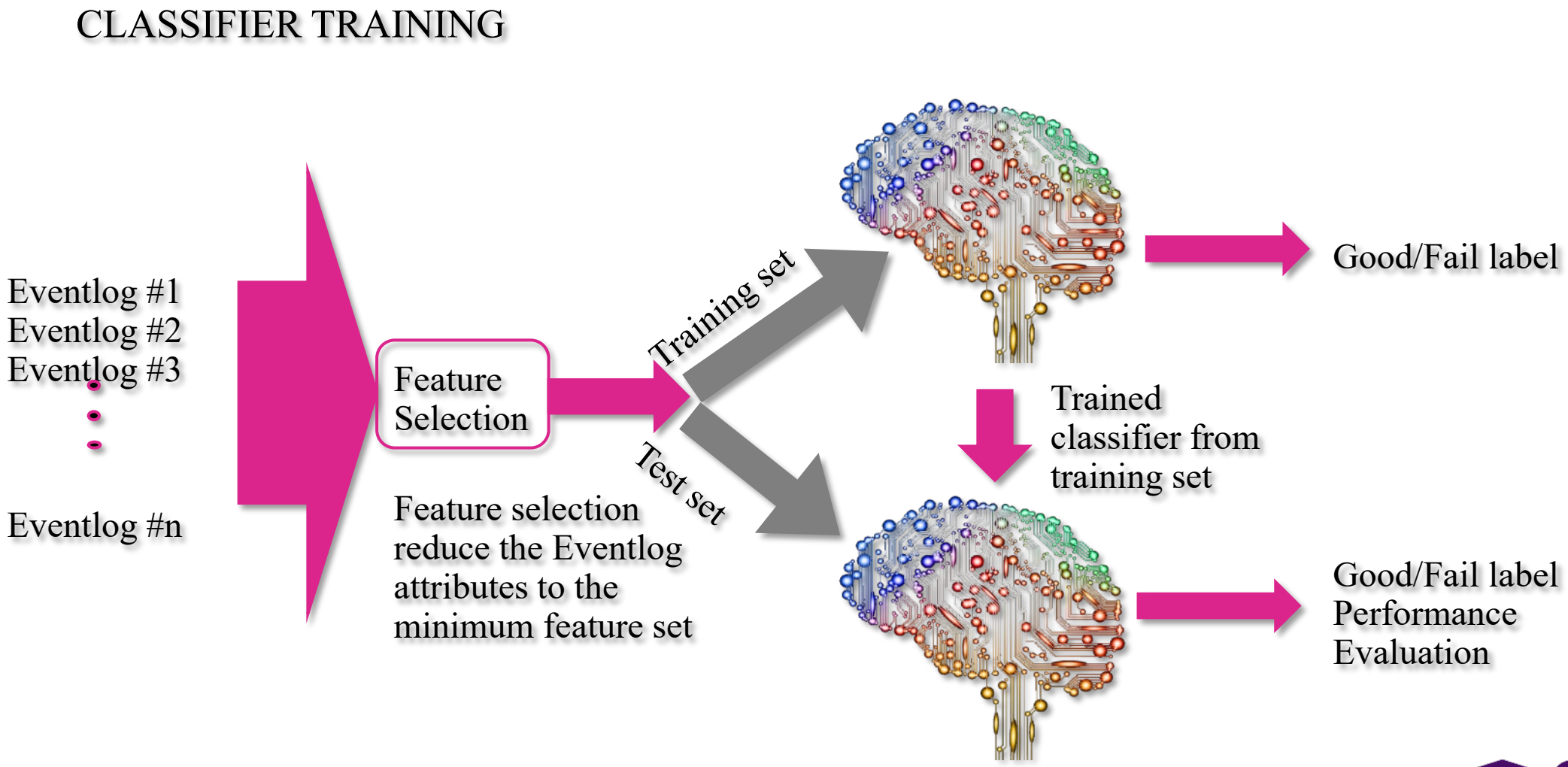
Dual Machine Learning (ML) Models

- The first ML model can predict the imminent component failure for preventive maintenance.
  → Storage device quality control

- The second ML model can forecast the quantity of long-term failure parts for supply chain inventory control.
  →Supply chain management

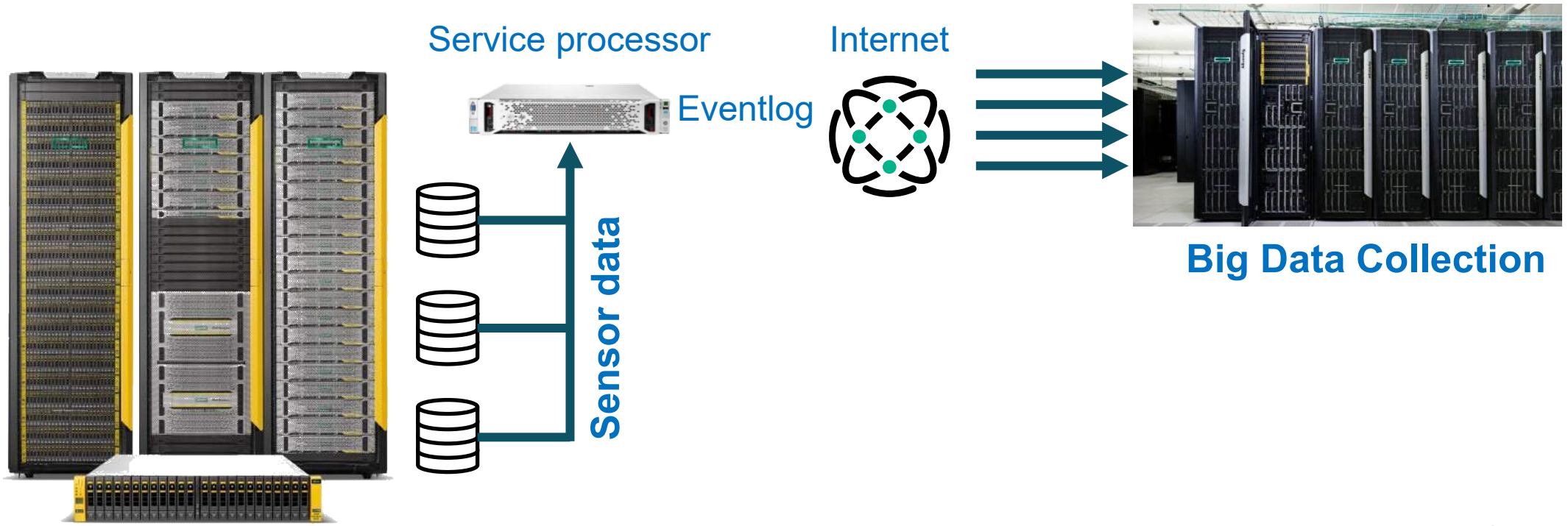
→ Both models require sensor data on electronic devices in a storage system.

Prediction model generation



Sensor Data collection

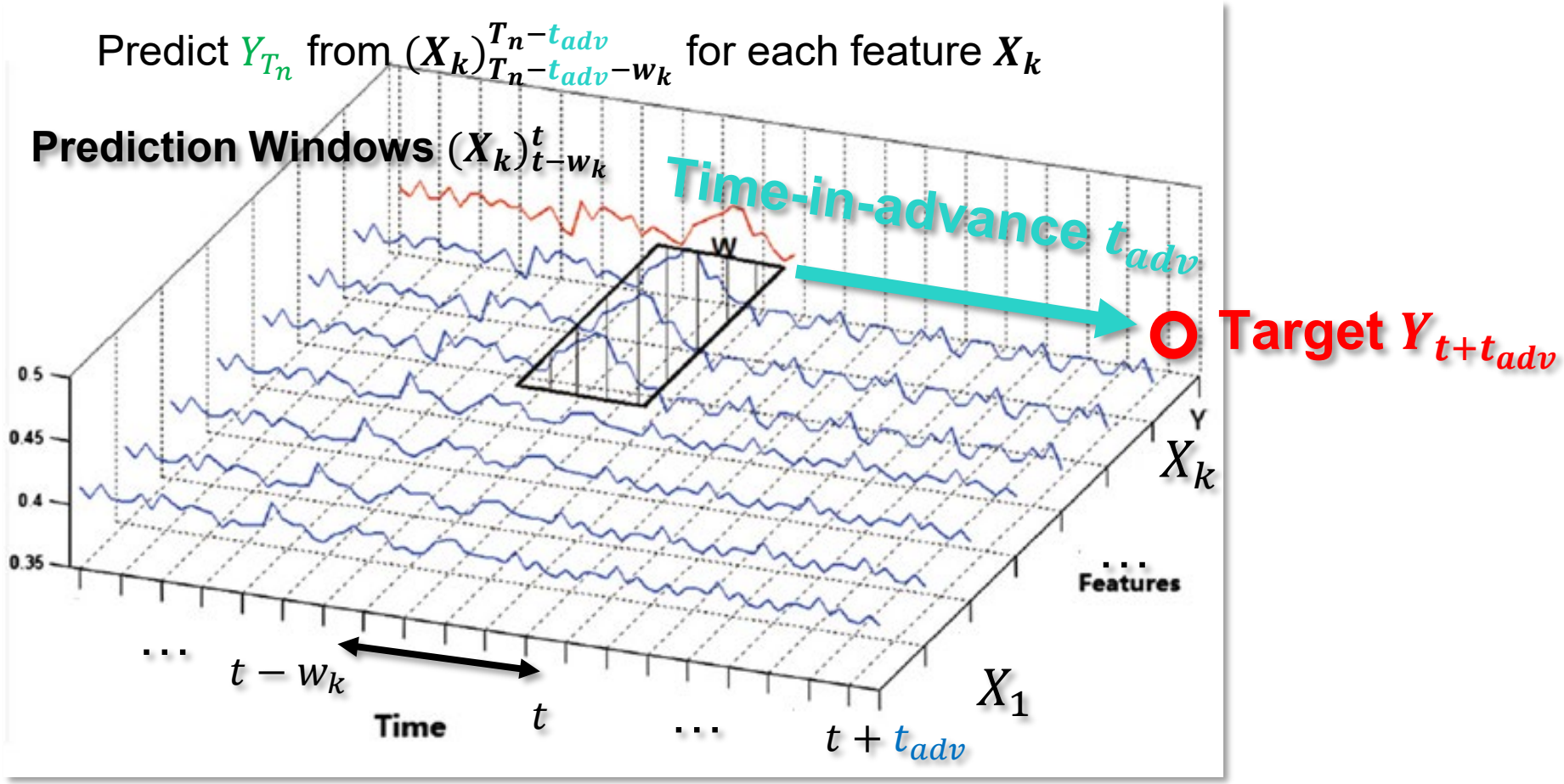
- Keep track of millions of electronic devices in the customer data center
- CallHome program collects the sensor data of the customer side.



Sensor data

- Time-series data
- All kinds of environmental activities (temperature, voltage,…)
- Errors values of electronic devices
- For Hard disk drive(HDD)
  - Read / Write correctable / uncorrectable errors
  - Read / Write I/O counts per drive per second
  - Provide the detailed diagnostic information about the various error conditions
- Network traffic by # of I/O counts

Casual Inference for feature selection



- 30% features reduction
- 15% accuracy improvement

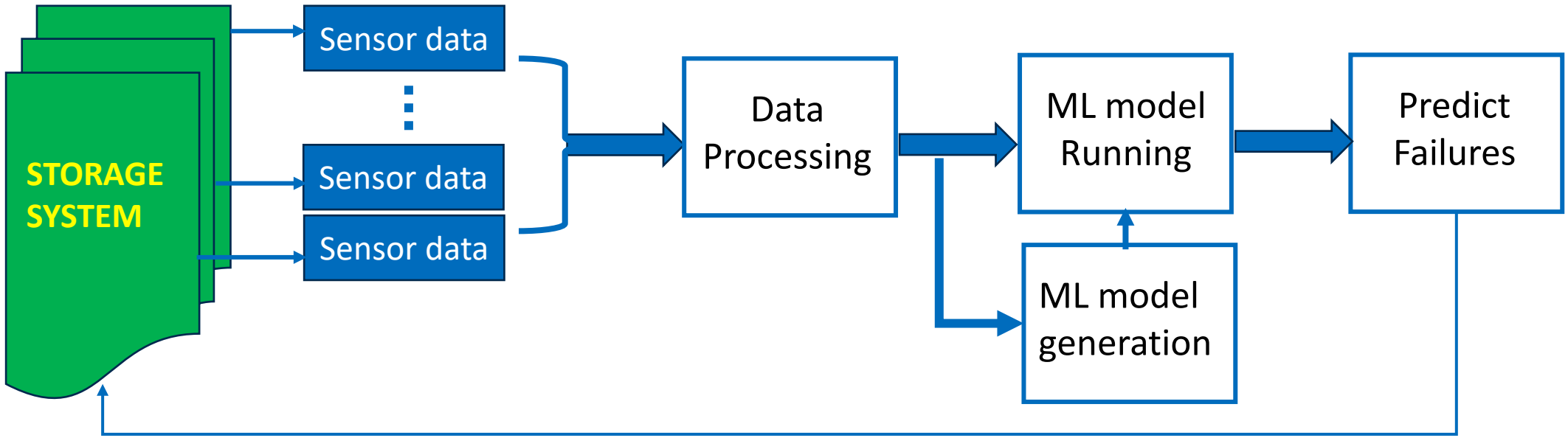
First ML model: Prediction of HDD failure

- The model predicts HDD as impending “fail” or “good” with HDD sensor data.
- The model is trained with the historical data of both fail and good HDD.
- eXtreme Gradient Boosting (XGBoost) used as ML model
  - Build an ensemble of decision trees sequentially
  - Each tree is trained to correct the errors of its predecessors

- Which HDD sensor data is included for the model generation?
  - XGBoost feature importance rank the sensor data

Machine Learning (ML)-driven predictive maintenance

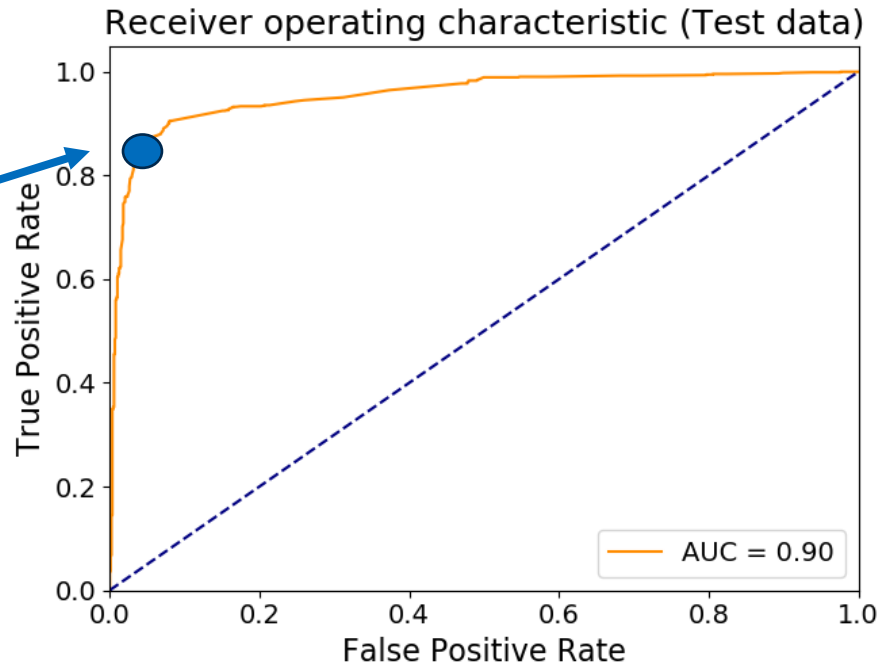
- Workflow



First ML model: Performance metric

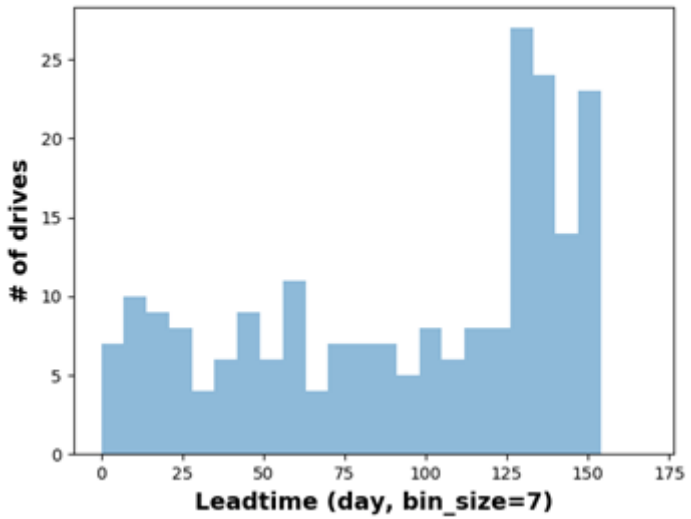
- ROC (Receiver operating characteristic) curve
  - TPR (True Positive Rate), FPR (False Positive Rate)
  - The orange curve is TPR vs FPR across all thresholds.
  - The closer to the upper left corner is the higher accuracy (TPR=1, FPR=0)

For example, TPR=0.86, FPR=0.03



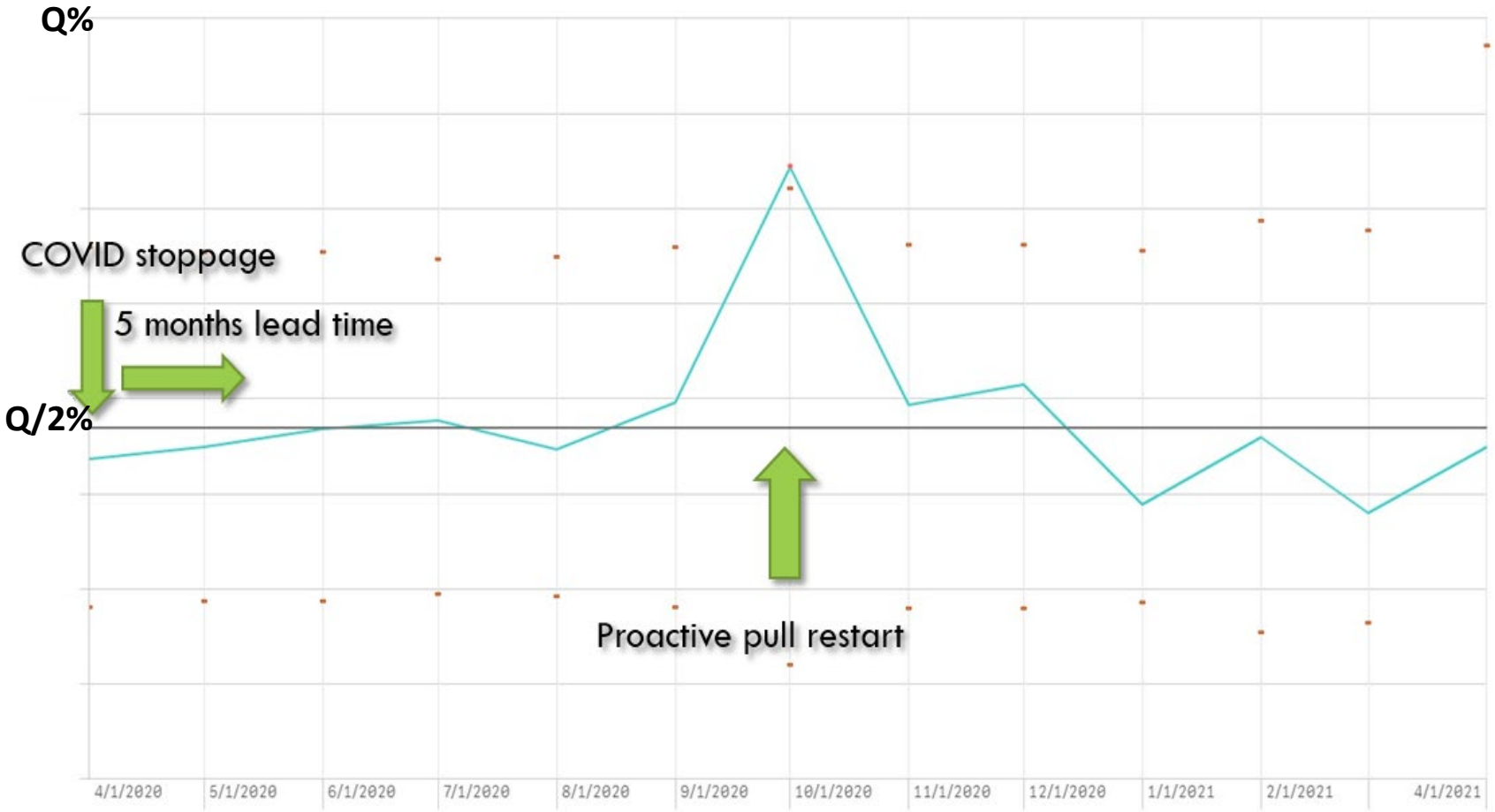
First ML model: Lead Time

- Lead time is the time from model detection to actual failure.
- From issuing the service tickets to replacing drives, the long lead time is required.
- Our model shows ~ 5months lead time.



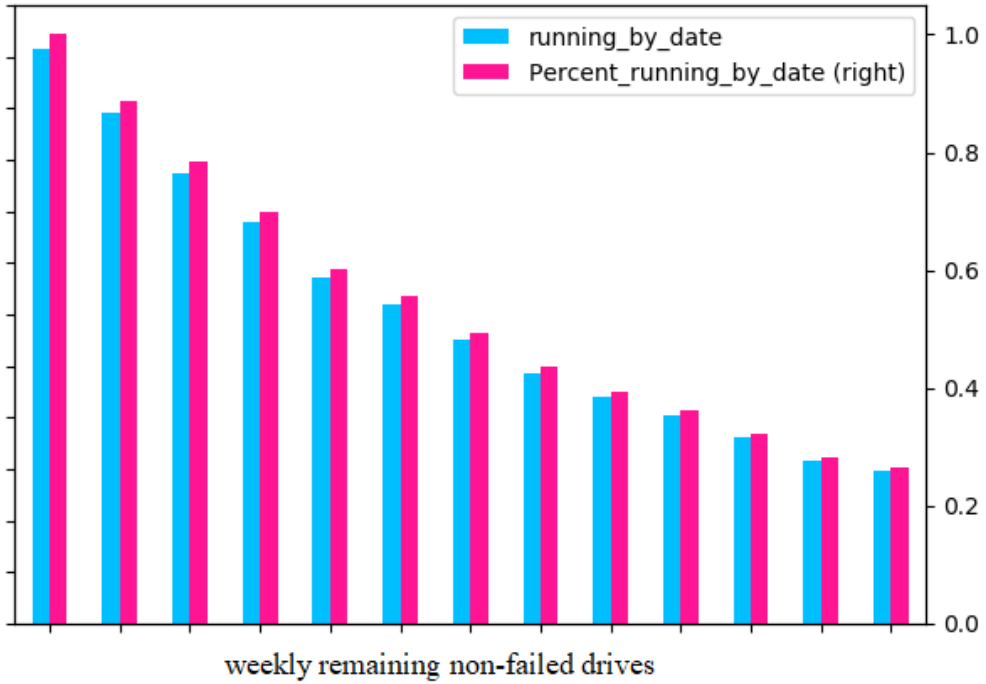
Covid shutdown / reopen

- Failure rate w/o and w/ proactive HDD removal, lead time ~5 months



The second ML model

- Supply chain quantity estimation model
  - Need a model to guarantee procurement time that acquires the new parts before failure
  - By using XGBoost model, count the residual unfailed drives



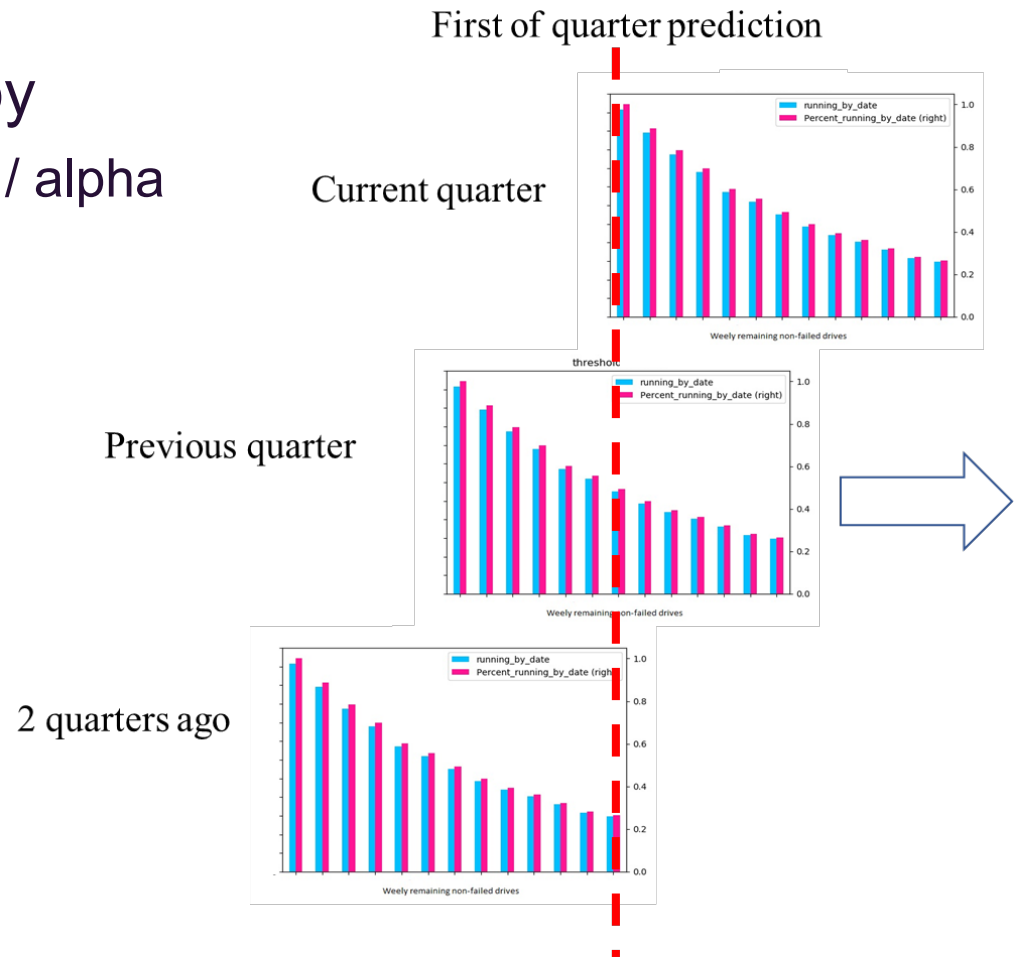
The second ML model

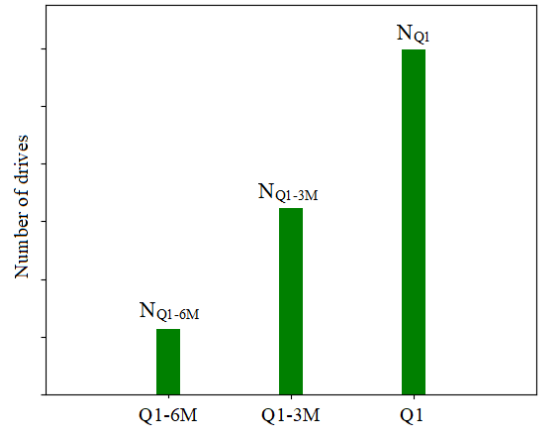
- Quarterly, # of failure demand is estimated by

  = ($N_{Q1\text{-}6M}$+ $N_{Q1\text{-}3M}$ + N_{Q1}) / alpha

  Alpha is derived by the regression and TPR

- Reduce the risk of the supply chain shortage

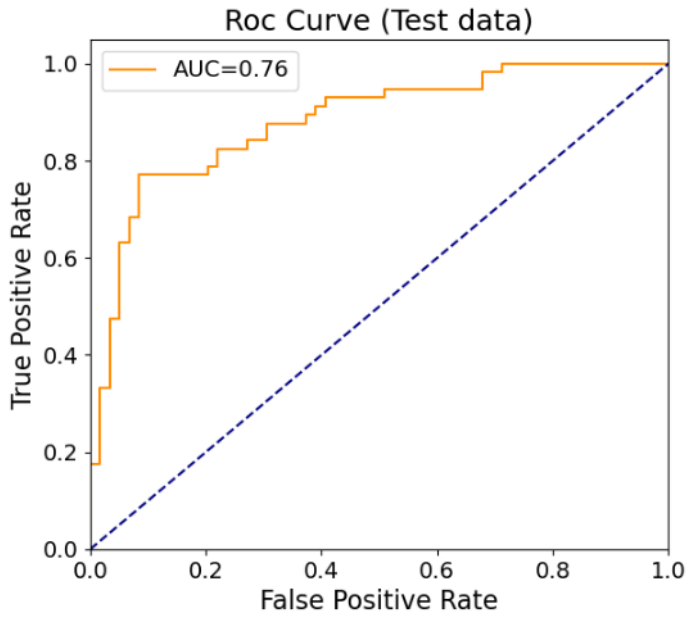




Q1 accuracy :96%
Q2 accuracy : 95%

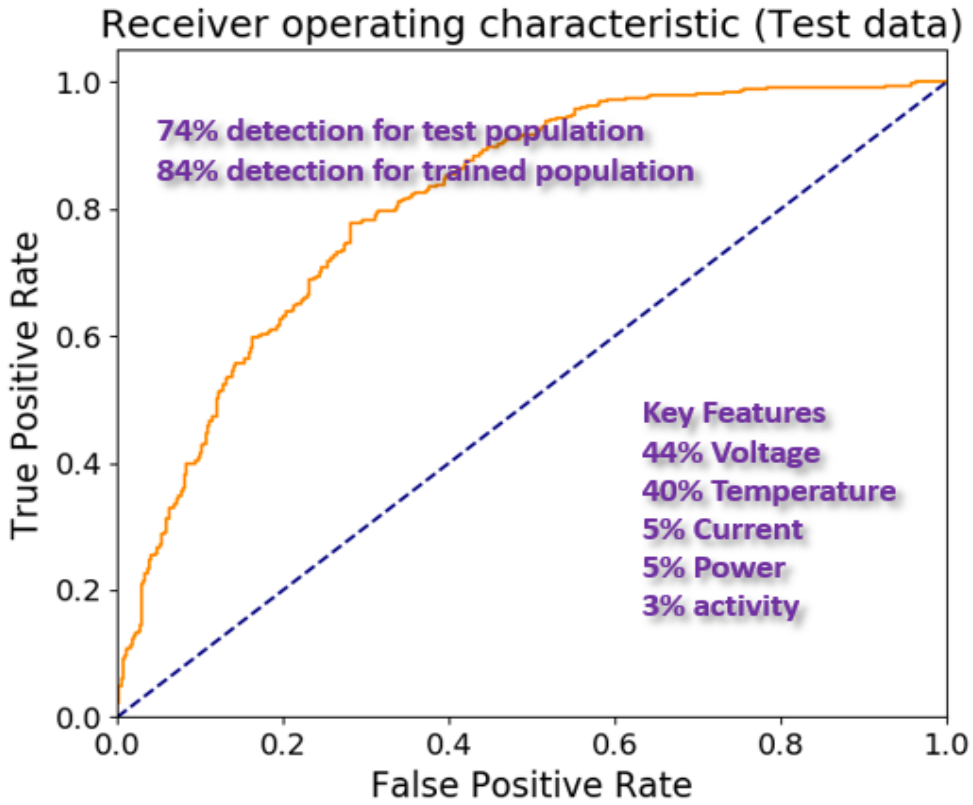
SSD Failure Prediction

- New sensor features model
- Early deployment success
  - Before deploy in a region : 80+ failures
  - After proactive removal :4 unanticipated failures
  - 98% in accuracy in failure quantity one quarter ahead of time

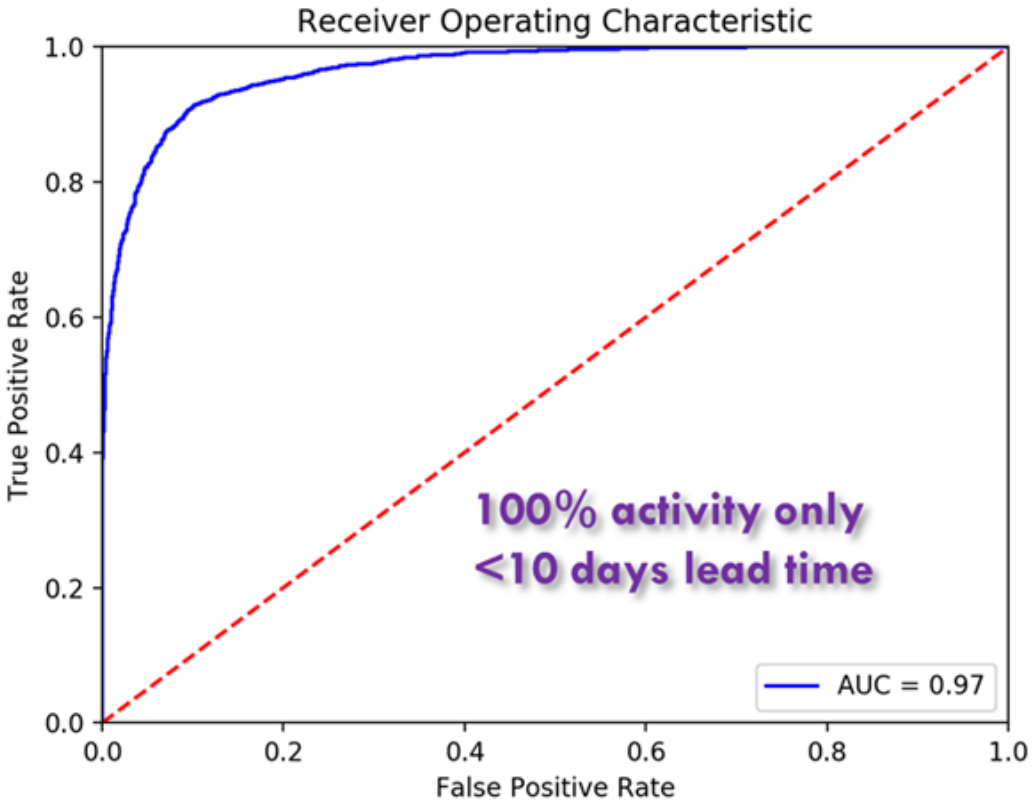


Voltage Regulator failure prediction

Long lead time prediction model
> 6 months before system failure



Detection of imminent fail in storage system



Conclusion

- Presented the predictive maintenance model of storage component failure
- The quarterly failure prediction model is developed by using the short-term failure prediction model.
- The quarterly failure model helps the supply chain management.

- The collection of historical sensor data is required to train the model.
- The selection of the sensor data related to the failure is the most critical step.
- The lead time of the model will decide if the model is deployable or not.
- The trade-off between accuracy and the lead time is unavoidable.

Please take a moment to rate this session.

Your feedback is important to us.

