Cloud Health:
A Forest of Hard Drives

Rob Johns

Metis Bootcamp

Summer 2019

The Cloud is Made of Hard Drives

==

In a Data Economy – Let Data Protect Itself

- Hard Drive failure huge cost risk to users and companies
- Data on 32K Hard Drives published by Blackblaze
- 120 days before failure
- Fewer drives – more storage

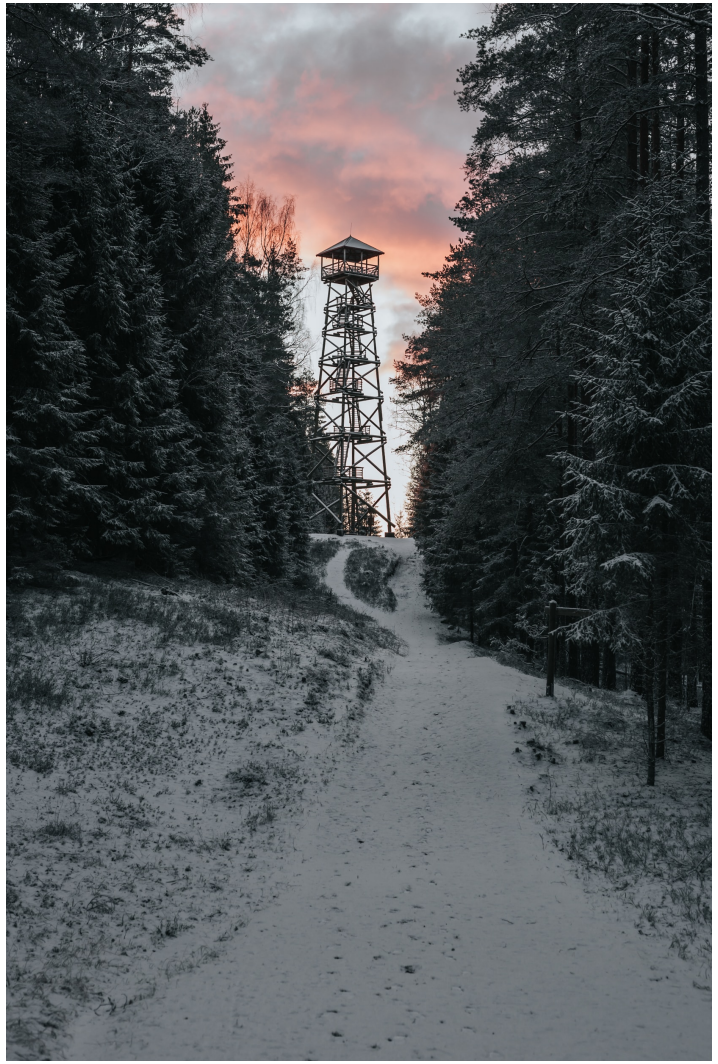
Random Forest Model can Preempt Fails

- Random Forest classification is fast and powerful for predicting failures
- Trained on 26K drives (4% failed)
- Prediction on 6.5K drives:
  - Correctly predicted 54% of fails (161 Drives)
  - Failure identified with precision of 87%
  - Incorrectly predicted 0.4% working drives will fail (24 Drives)
  - Proper function identified with precision of 98%

Hard drives are S.M.A.R.T.

- Most Important Errors:
  - Read errors
  - Spindle start/stops
  - Logical block addressing reads
  - Drive usage hours
  - Temperature deviance from optimal
  - Head engage/disengage

Business Applications: Passive Forest Surveillance



- Typical server farms have a lot of drives
- This model is fast – could run hourly on a single computer for a whole farm
- 2 days warning will help:
  - Model still predicts 35% of fails
  - Model still predicts 99% proper function

Conclusion and outlook

- Random forest model can help avoid data loss by identifying failing drives without adding much overhead
- Models can be made predictive to allow human testing interaction with drives with days notice
- Model suggests drive on/off cycling, energy saving modes and heat management key to drive health