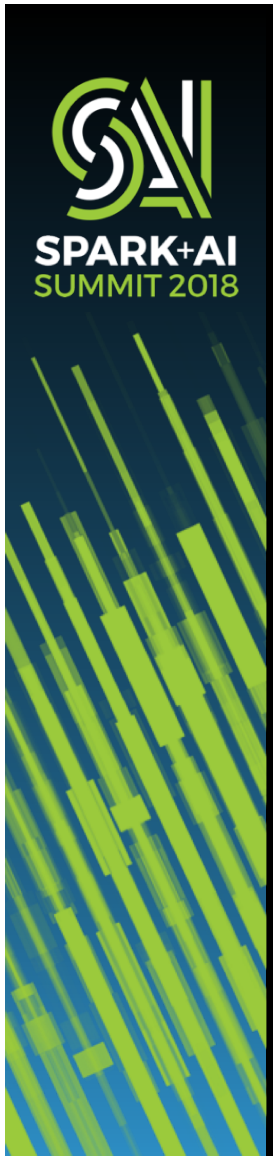


How Azure Databricks helped make IoT Analytics a reality

Prasad Chandravihar, Lennox International
Janath Manohararaj, Lennox International

#Ent7SAIS

Agenda

- About Lennox
- Problem Statement
- How did we start
- Challenges
- Machine Learning using Databricks
- Wrap – up
- Q&A

About Lennox

- Lennox International Inc. is an intercontinental provider of climate control products for the heating, ventilation, air conditioning, and refrigeration markets.
- Founded in 1895, in Marshalltown, Iowa, by Dave Lennox.
- Three core businesses: Residential Heating and Cooling, Commercial Heating and Cooling, and Refrigeration.
- The company headquarters are in Richardson, Texas, near Dallas.

Problem Statement

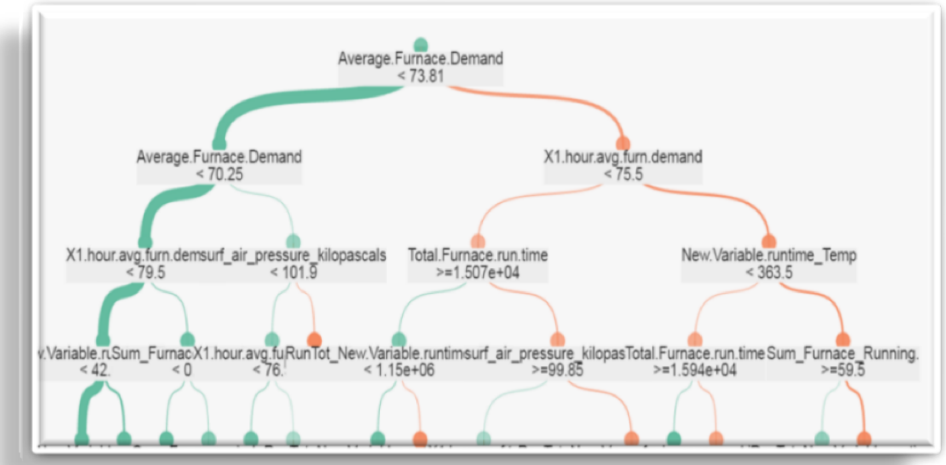
Particular type of mechanical switch trips pretty often and causes discomfort to the users

Use data sets captured from the IoT devices to analyze

Build ML models to predict the trips and also detect patterns/influencing variables

How did we start

- Initial Model was developed in desktop tools with about 15 devices
- Accuracy levels – Recall : 65%, Specificity : 80%

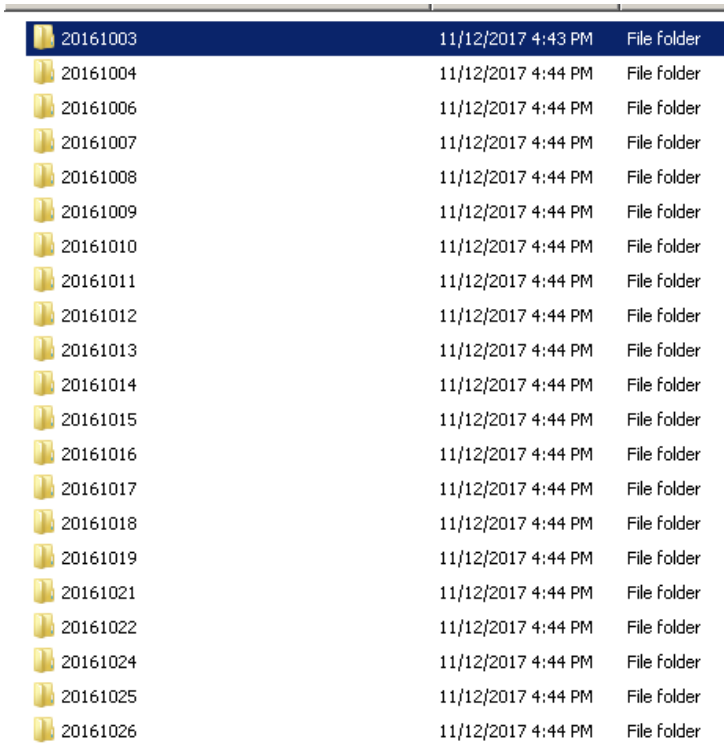


Challenges

- Consuming the entire data set
  - Data orchestration – 10 Billion rows
- Assembling different skill sets
  - Team structure, working model
- Right Computing platform – Cloud ?
  - collaboration among Data Scientists & Engineers

Small file problem

- Traverse through 7 million directory paths to gather data



20161003	11/12/2017 4:43 PM	File folder
20161004	11/12/2017 4:44 PM	File folder
20161006	11/12/2017 4:44 PM	File folder
20161007	11/12/2017 4:44 PM	File folder
20161008	11/12/2017 4:44 PM	File folder
20161009	11/12/2017 4:44 PM	File folder
20161010	11/12/2017 4:44 PM	File folder
20161011	11/12/2017 4:44 PM	File folder
20161012	11/12/2017 4:44 PM	File folder
20161013	11/12/2017 4:44 PM	File folder
20161014	11/12/2017 4:44 PM	File folder
20161015	11/12/2017 4:44 PM	File folder
20161016	11/12/2017 4:44 PM	File folder
20161017	11/12/2017 4:44 PM	File folder
20161018	11/12/2017 4:44 PM	File folder
20161019	11/12/2017 4:44 PM	File folder
20161021	11/12/2017 4:44 PM	File folder
20161022	11/12/2017 4:44 PM	File folder
20161024	11/12/2017 4:44 PM	File folder
20161025	11/12/2017 4:44 PM	File folder
20161026	11/12/2017 4:44 PM	File folder

Solution:

Developed a batch program which would gather data from each file in each destination and append to a table

Machine Learning Core team

Data Engineering

- Helps on identifying the right data sets
- Helps in data orchestration

Data Science

- Helps on getting the data ready for the model (parts of ETL)
- Explore different data science models

Platform Eng.

- Analyzes the feasibility to productize

Eng. SME

- Feedback on the outputs
- Helps in making sure the analysis is progressing in the right direction

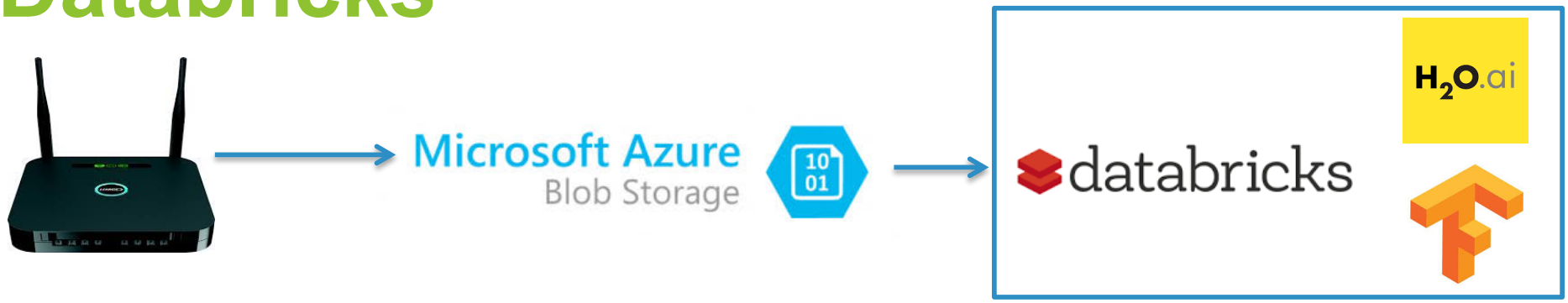
Big Data Cloud Architecture - helps the team on the best use of the tools and resources

- Constantly works with Data science and Engineering members during data orchestration

Why Azure Databricks ?

- Unified platform for Machine Learning & Data engineering
- Provides collaboration between team members
- Spark with minimal tuning
- Average time to start a cluster – 4 mins
- Automatic scaling – very important as different jobs are different sizes
- Ability to integrate sparkling water, H20 driverless AI and Tensorflow through GUI
- Ability to integrate PowerBI

Machine Learning using Azure Databricks



- Sends data every minute

- Equipment information
- Weather information
- Demographic information

- ELT pipeline
- Training the ML model
- Validation and scoring

70% of the time – structuring and preparing the data, creating labels etc..

Machine Learning using Azure Databricks

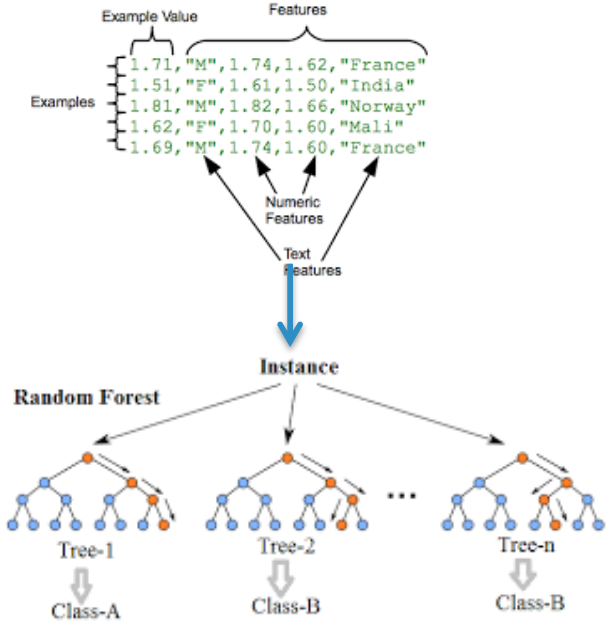
Class Imbalance:

We had way more 0's than 1's as labels , for ex: for every 590M 0's we had 5K 1's.

We then tried 3 different approaches:

- Over sample the 1's (SMOTE, duplicating rows etc..)
- Under sample the 0's (relatively small set of 0's which represent the population)
- Multiple samples of 0's and the same set of 1's
  - each set of 0's & 1's will have its own model
  - Created the ensemble of models

Machine Learning using Azure Databricks



First sample of 0's & 1's with the assembler of features

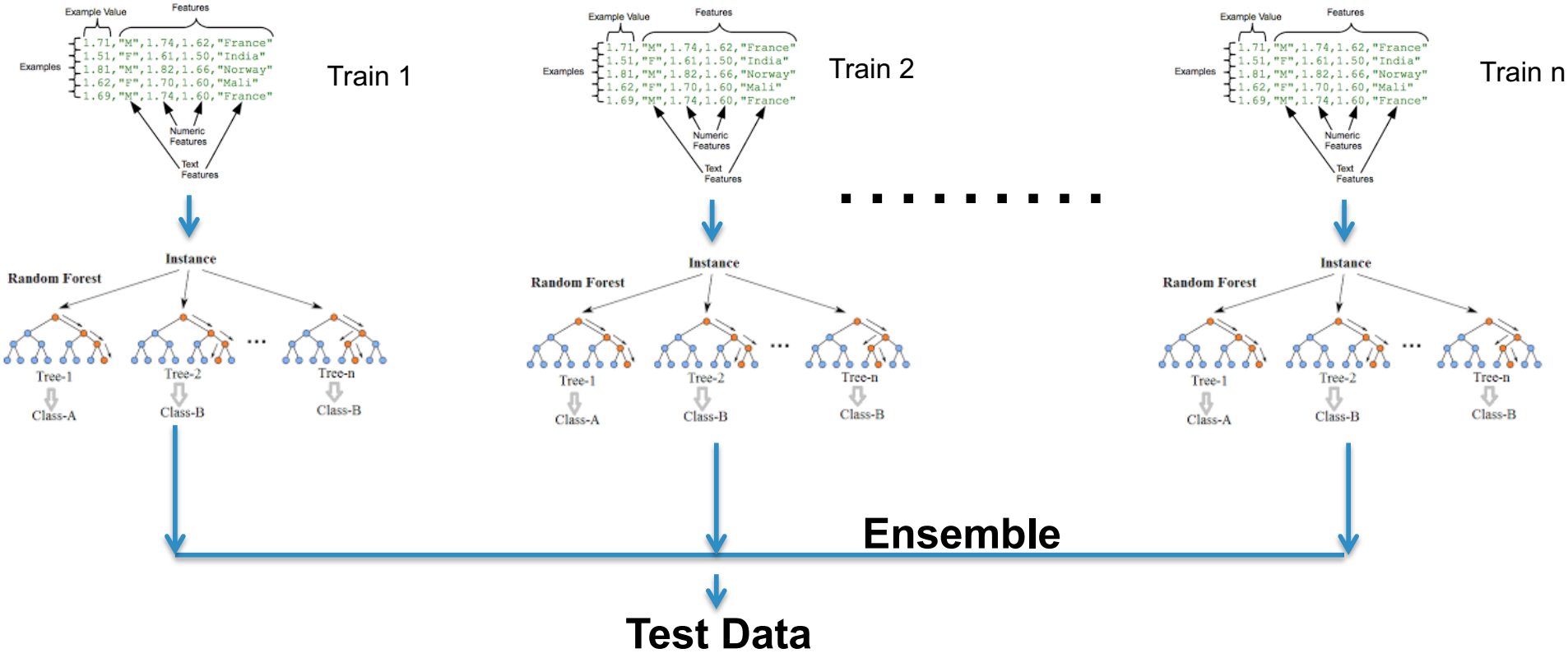
#Bins, #trees, #depth and threshold tuned using an exhaustive grid search

Sample output from Grid search

```
depth = 10 trees = 100 bins = 100 Area under PR = 0.06771043665922438
depth = 10 trees = 100 bins = 120 Area under PR = 0.07241627620184914
depth = 10 trees = 100 bins = 150 Area under PR = 0.07182619204294148
depth = 10 trees = 100 bins = 200 Area under PR = 0.06734260472997455
depth = 10 trees = 150 bins = 100 Area under PR = 0.06562223133217317
depth = 10 trees = 150 bins = 120 Area under PR = 0.06807527695347144
depth = 10 trees = 150 bins = 150 Area under PR = 0.06795915416548445
depth = 10 trees = 150 bins = 200 Area under PR = 0.07104517441446621
depth = 10 trees = 200 bins = 100 Area under PR = 0.06579687173976631
depth = 10 trees = 200 bins = 120 Area under PR = 0.07384870933349169
depth = 10 trees = 200 bins = 150 Area under PR = 0.07503096420318
depth = 10 trees = 200 bins = 200 Area under PR = 0.06573181561929348
depth = 10 trees = 300 bins = 100 Area under PR = 0.06965294118857658
depth = 10 trees = 300 bins = 120 Area under PR = 0.06941528348545122
depth = 10 trees = 300 bins = 150 Area under PR = 0.06652078632433465
depth = 10 trees = 300 bins = 200 Area under PR = 0.06816366084407827
depth = 15 trees = 100 bins = 100 Area under PR = 0.1187958164849981
depth = 15 trees = 100 bins = 120 Area under PR = 0.11801598980930454
depth = 15 trees = 100 bins = 150 Area under PR = 0.1234935252087955
depth = 15 trees = 100 bins = 200 Area under PR = 0.11392714942776622
depth = 15 trees = 150 bins = 100 Area under PR = 0.13368842649032905
depth = 15 trees = 150 bins = 120 Area under PR = 0.12101158733224104
depth = 15 trees = 150 bins = 150 Area under PR = 0.11547666200057739
depth = 15 trees = 150 bins = 200 Area under PR = 0.11863689140472232
```

Model was tuned for the best Area under Precision Recall

Machine Learning using Azure Databricks



Machine Learning using Azure Databricks

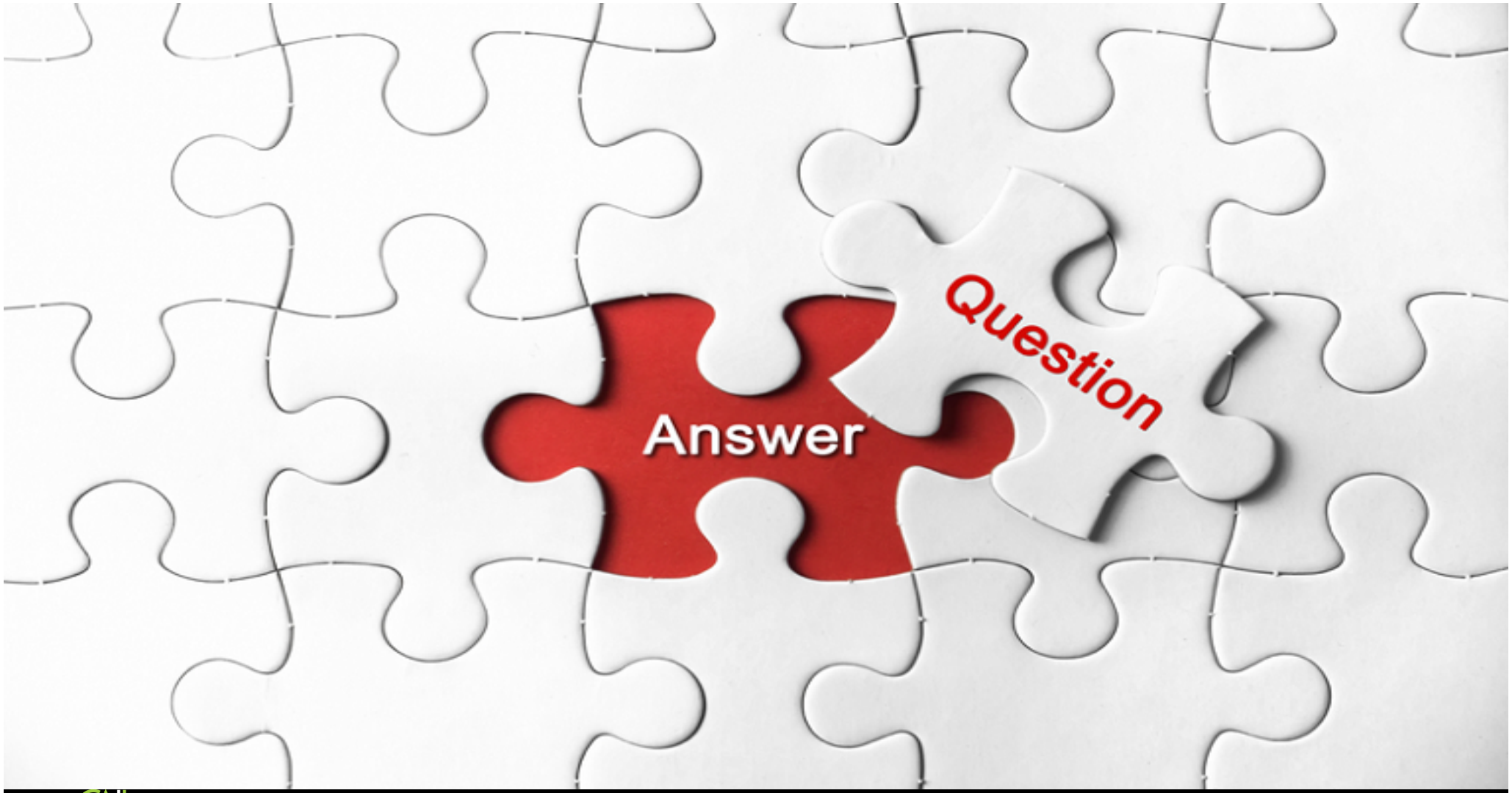
- Right from data engineering to ML models was done in one single notebook
- Coding language - PySpark
- Each ML model was tuned to get the best hyper parameters using an automated grid search
- The ensemble model helped us reduce a lot of false positives

Wrap – up

Journey to ML	15 devices	500 devices	25K devices
Accuracy	TP - 65%	TP - 80%	**TP – 84.6%**
	TN - 80%	TN - 85%	**TN – 99.5%**
Data Volume	2 Million	100 Million	**10 Billion**
Tools	Desktop tools	Spark ML,H2O	Azure Databricks
Cloud Services Provider	n/a	Microsoft	Microsoft
ML models used	Decision Tree	Multiple (Gradient Boosted Trees, Random Forest, etc)	Multiple (Gradient Boosted Trees, Random Forest, etc)
Time to run ML model	6 hours	30 mins	**50 mins**

Wrap – up

- Build a cross functional team to execute machine learning projects
- In most of projects 70% of the time is spent on cleansing and transforming the data set
- Give a lot of focus into engineering features
- Explore sparkling water (H20 on databricks) gives a lot of auto ML options
- Platform which lets team members collaborate and develop the project end to end


Answer
Question