Welcome to the webinar!

Rladies Milan

Tuesday 28th July- 18:30

Root cause analysis using frequent pattern discovery

Golnazsadat Zargarian
Data Scientist
Altran Italia

Data science

Return of experience

High demand for data science!

Pandemic changed the online behaviour

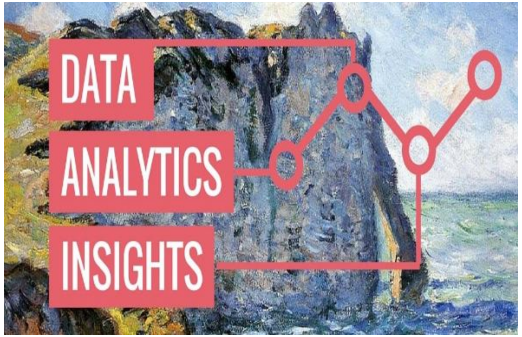
- **Italy**: - search for mailing services peaked at 114%
  - clicks to call to these businesses reached 198%
  - peak for online groceries

Data is the new oil ☺

- Data is generated quickly and for many companies there is a need to make challenging decisions.

And data science is shaping the future more than ever!

Who is a data scientist?



Data

Collect data, pure information. Real data is messy and often invaluable.

Analytics

Analytics is there to make sense out of data. It could discover patterns and trends.

Insights

Interpret analytics results and provide valuable insights.

So, where to start?

How to grow professional network?

Take initiatives

- Engage with like-minded people
- Follow top-voices
- Attend in meet-ups

How to level up the skills needed?

Take Courses

- Soft skills are quite important
- Educational background

How do to create a portfolio?

Build projects

- Work on real world data
- Apply different methods and compare


Altran
Career opportunities

Who are we?

altran
Part of Capgemini

We are the undisputed global leader in Engineering and R&D services with a local footprint in **30+ countries** and **45000+ employees**.

Aeronautics
AIRBUS · SAFRAN · ROLLS ROYCE

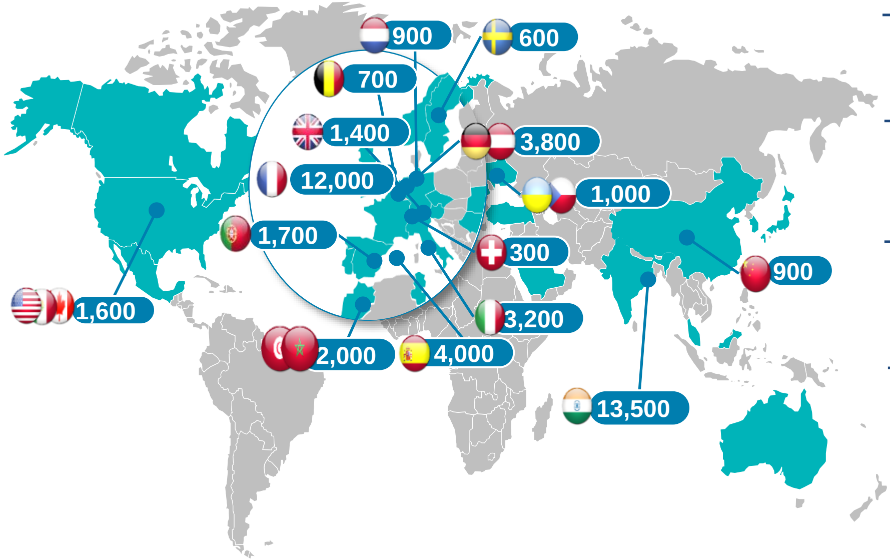
Space, Defense & Naval
AIRBUS DEFENCE & SPACE · THALES · DASSAULT AVIATION

Rail, Infrastructure & Transport
ALSTOM · BOMBARDIER · SNCF

Energy
ENGIE · EDF · GE

Industrial & Consumer
Schneider Electric · AIR LIQUIDE · Whirlpool

Life Sciences
SANOFI · gsk · Johnson&Johnson



Automotive
PSA GROUPE · BMW · BOSCH

Finance & Public Sector
BNP PARIBAS · HSBC

Semiconductor & Electronics
Qualcomm · AMD · ASML

Software & Internet
IBM · amadeus · Microsoft

Communications
vodafone · AT&T · CISCO · NOKIA

Altran's World Class Center for Analytics

EXPERIENCE

Our approach to data science has continually refined over 40 years of work on 1000s of projects for 100s of clients across a range of complex domains and industries.

PURPOSE

Our focus is on delivering business value & impact from AI and data science. Our culture, processes and people are all aligned to our mission.

SIZE

Over 350 data science & AI experts provide the range of talents needed for successful digital initiatives

REACH

US and EU offices provide the dedicated local resources valued by our clients alongside on-demand access to our global talent pool.

DATA MANAGEMENT
- Data Warehouse Design and Maintenance
- Master Data Management
- Data Quality

ETL
- ETL Procedure Design and Maintenance
- ETL Robustness Assessment

REPORTING
- Reports Design
- Dashboard Design
- Self Reporting

BIG DATA ARCHITECTURE
- Big Data Architecture Design and Sizing
- Distributions Selection and Configuration
- Full Text Server Engine

DISTRIBUTED CALCULATION
- Spark Architecture Configuration and Design
- Distributed Calculation Optimization

REAL TIME ANALYSIS
- Real Time Analysis Strategy Definition
- Real Time Component Design and Maintenance

MACHINE LEARNING
- Predictive Algorithms
- Deep Learning

OPTIMIZATION ALGORITHMS
- Operative Research Models
- Dynamic Time Warping
- Iterative Algorithms for linear and not linear systems

USE CASES DEFINITION
- Data and Features Exploration
- Feasibility Study
- Analytics Roadmap

Data intelligence use cases

Anomaly detection

Using historical video streaming data to detect anomalies.

Real Time Data Prediction

Predicting the real time behaviour of data based on different historical KPIs.

Anomaly Detection Engine

Predicting anomalies based on the prediction of data in real time.

Pattern Discovery

Real world business problem

Scope and vision

What is the Goal?

- The project is done for TIM in collaboration with Polito
- Making sense out of thousands of events/alarm logs, generated in the 3G/4G base stations per day
- Automatic extraction of situations
- Identify possible future anomalies

Why the problem is challenging?

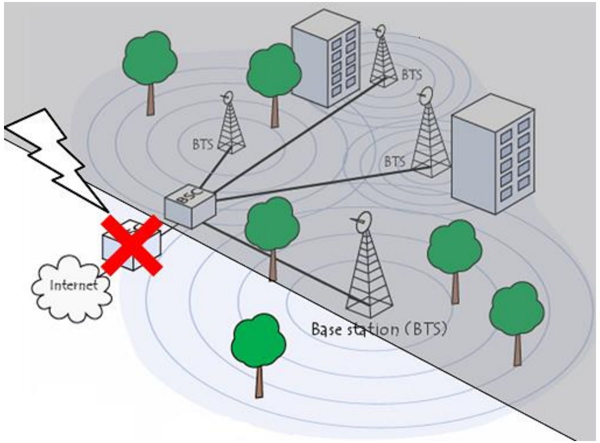
- Manual analysis is time consuming
- Dataset is heterogeneous and large

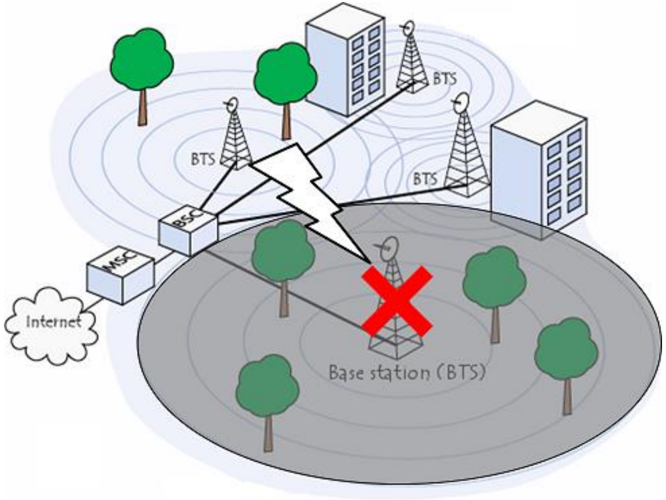
How to achieve the goal?

- Analyzing most important KPIs
- Understand the root cause of network failures by mining frequent patterns

Possible failures in mobile networks

Mobile switching center (MSC) failure





Base transceiver station (BTS) failure

Model Lifecycle: from start to end

1 Data Gathering and Preparation

- Historical Datasets of alarms
- Data Cleaning
- A look into Most Important Metrics

2 Data Characterization

- Understanding the data more!
- Data Distributions

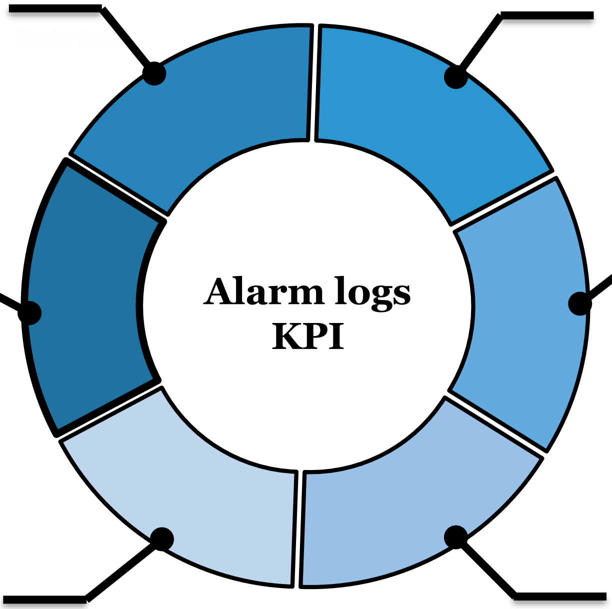
3 Choosing an ML model

- Root-cause detection
- Sequential pattern mining
- Association rule extraction

4 Evaluation and Parameter Tuning

- Checking the Error Distribution
- Validating the model
- Checking if patterns are correct

Effective KPIs



Network ID: Device ID which generated the alarm

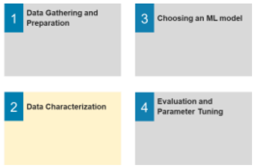
Timestamp: Start and end of alarm log

Alarm severity: Cirtical, warning, etc

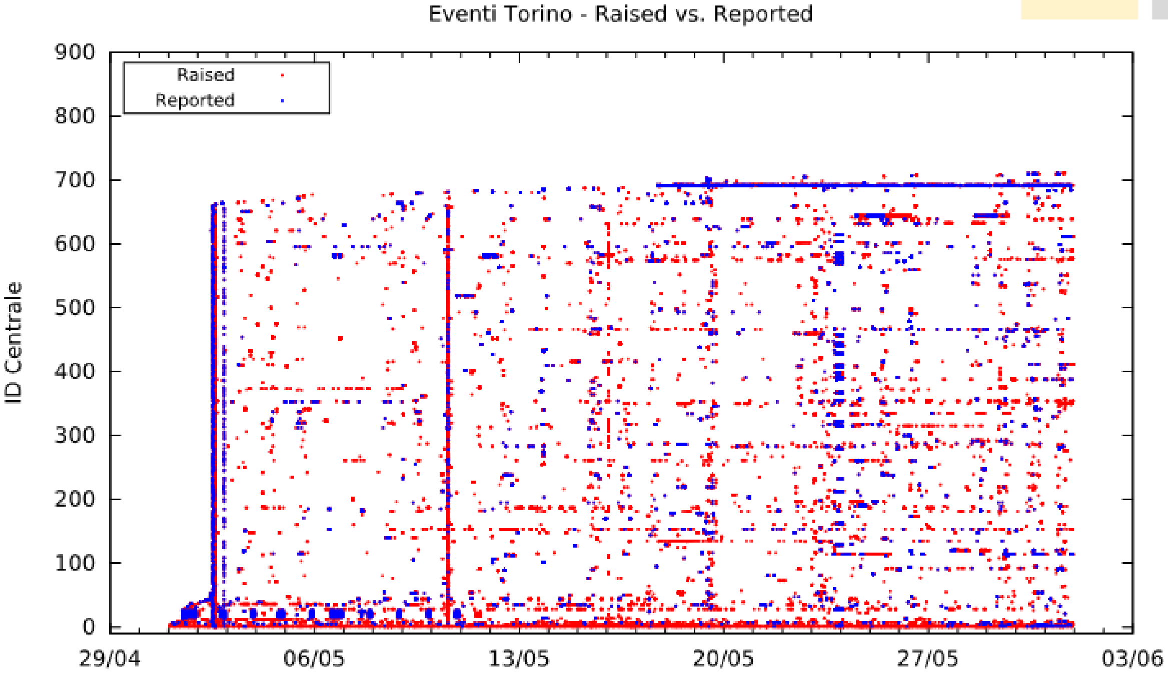
Alarm type: communication, equipment, processing, etc

Probable Cause: Primary cause of the alarm different from vendor to vendor

Site coordinates: Longitude and latitude of device

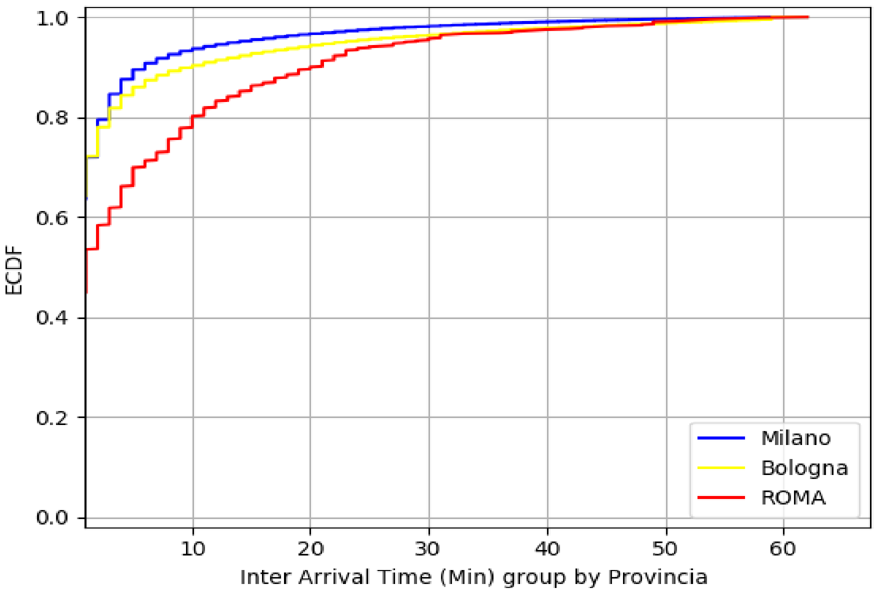


Data Characterization



Temporal evolution of events generated by network devices of Turin province in May.
Reported alarms were reported to the domain expert.

Data Characterization



CDF of Inter-arrival Times Grouped by Province

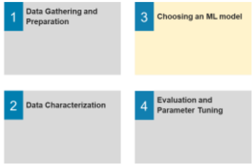
More than 40% of alarms are **co-occurrence**!

Market basket analysis

ID	Items
1	Bread, Milk
2	Bread, Pizza, Beer, Eggs
3	Milk, Pizza, Beer, Cola
4	Bread, Milk, Beer, Eggs, Pizza
5	Milk, Cola

- What are the elements that appear together frequently?
- What are the highest conditional probabilities?
- Pizza $\Rightarrow$ Beer

Market basket analysis



TID	Bread	Milk	Beer	Eggs	Pizza	Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	0	1	1
4	1	1	1	1	1	0
5	0	1	0	0	0	1

- To apply frequent pattern mining algorithms, we need a binary representation of data.
- Top possible algorithms: FP-Growth, A-priori

Rules: Metrics of importance

- Rule $x \Rightarrow y$

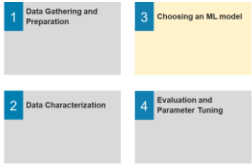
Support: How frequently the item appears in dataset

Confidence: How often the rule is found to be true

Lift: Interprets the significance of the rule

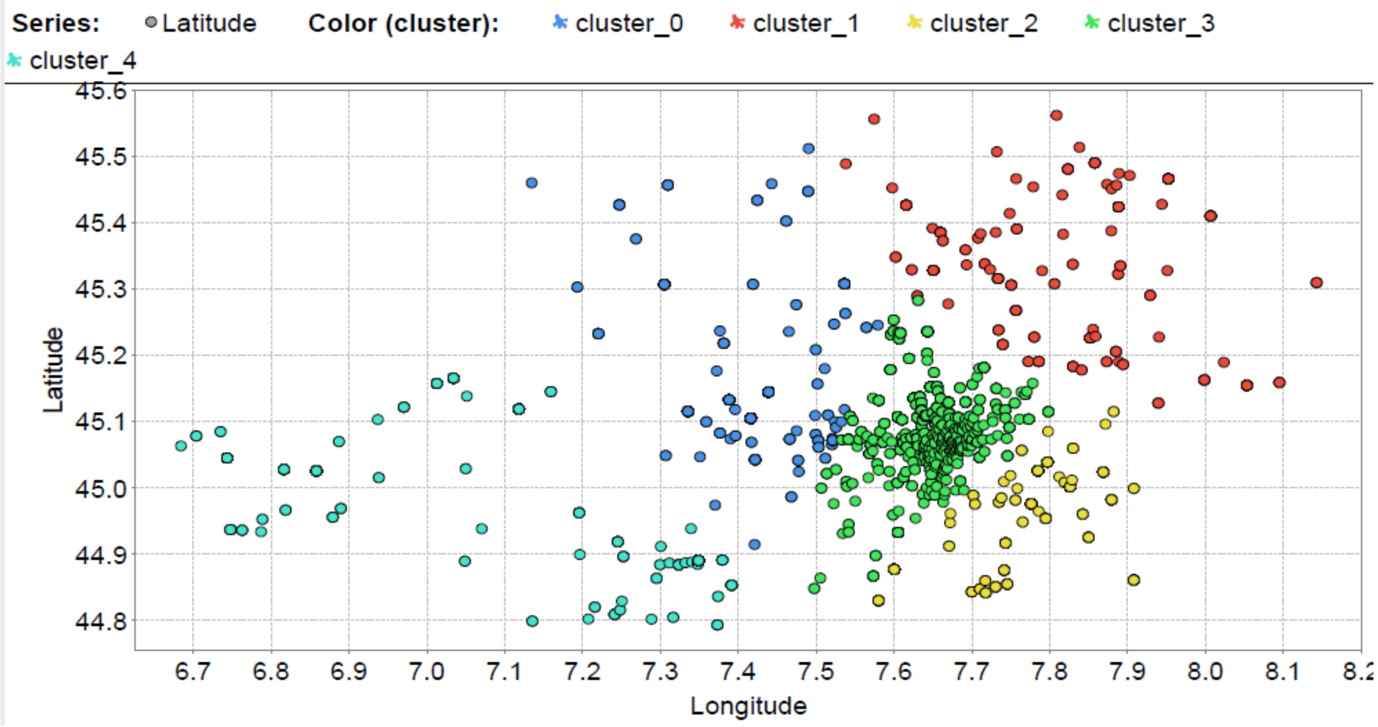
$$confidence(x \Rightarrow y) = \frac{support(x \cup y)}{support(x)}$$

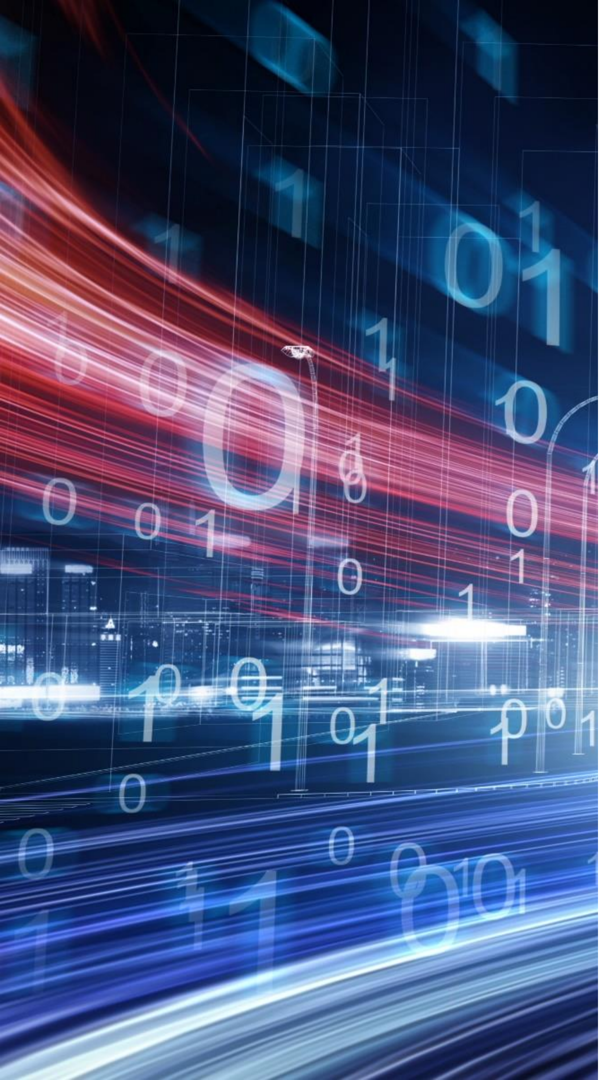
$$lift(x \Rightarrow y) = \frac{confidence(x \Rightarrow y)}{support(y)} = \frac{support(x \cup y)}{support(x) \cdot \ support(y)}$$

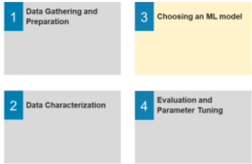


Spatial clustering of network devices

K-means clustering method, with k chosen based on the size of region (Turin) and RNCs.







Transaction Definition

Transaction ID (two hour window)	Device 1	Device 2	...	Device 320	Device 321
1	1	0	...	0	1
2	1	1	...	0	0

Device 1 raised alarms in the first time window period

Device 321 raised no alarms in the second time window

- Consider each cluster as a transaction matrix (Turin=5)
- Each transaction is a two hour window (372 bins for one month).
- Each item is a device ID (321 devices for Turin).
- Binary representation

R to rescue!

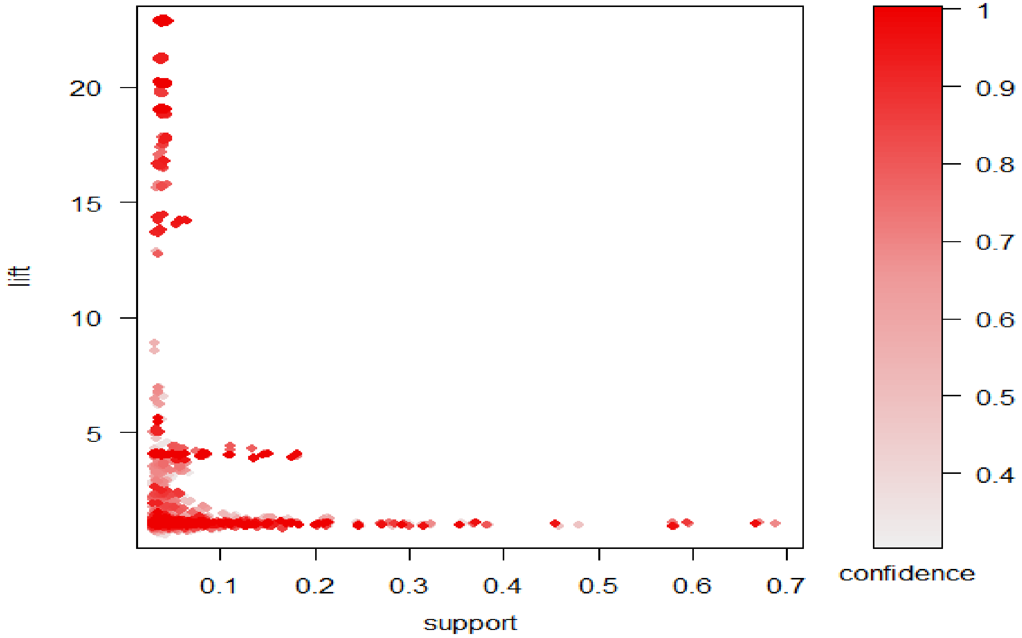
- Looking for a huge amount of patterns is not an easy task.
- R has a nice package for **visualizing association rules:**
  - Easy to code
  - Easy to understand the results
  - Works with different kind of frequent pattern mining algorithm

- **arulesViz** is an interactive package for visualization of association rules and frequent itemsets with R.

- Read more here:
- https://journal.r-project.org/archive/2017/RJ-2017-047/RJ-2017-047.pdf
- https://cran.r-project.org/web/packages/arulesViz/vignettes/arulesViz.pdf

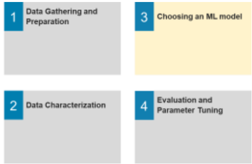
Scatter plot of rules

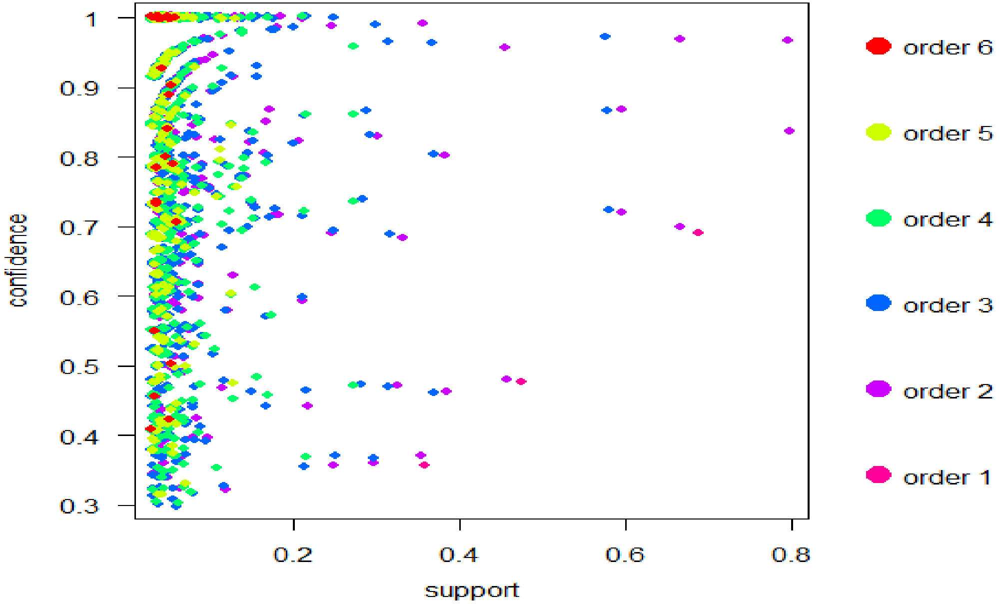
1 Data Gathering and Preparation
3 Choosing an ML model
2 Data Characterization
4 Evaluation and Parameter Tuning



- We select the most interesting rules based on lift, confidence and support.
- The rules that have a high value of lift have a support less than 10%.

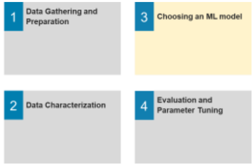
Two-key plot of rules





- Support and order have a strong inverse relationship
- Rules with many items involved are not that common!

Pattern example



- The rule involves 10 devices which makes it more interesting!

LHS: (33)

UBTSTO27F,
UBTSTO08E,
UBTSTO384

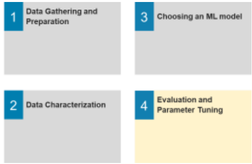
RHS: (32)

UBTSTO0B7,
UBTSTO14A,
8BTSTO384,
1BTSTO0B7,
8BTSTO0B6,
1BTSTO156,
1BTSTO00D]

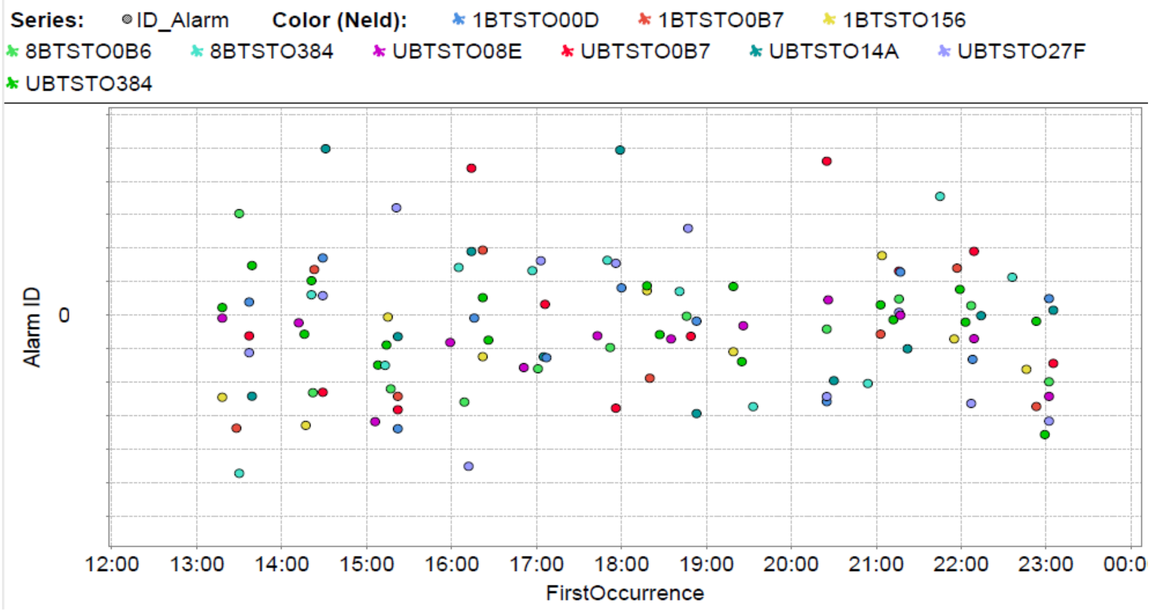
- Confidence: 0.97
- Lift: 11.15
- Support: 0.86

Location of devices

- Devices are located very close to each other. The rule confirms spatial correlation.

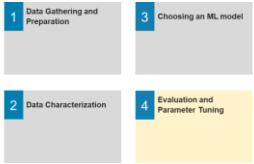


A closer look



- Strong correlation is observed among devices even at 20 minutes intervals. The rule confirms temporal correlation.

Conclusions and validations



- We aid the network operators, simplifying network management
- **Important**: we use different definitions of items and transactions
- Rule mining solutions, identifying rules significance and generalization

Feedback from TIM network maintenance team:

✓ Rules found were already manually registered in their system.

✓ TIM is now storing the list of pattern we created, presenting them together with other metadata.

Thank You!

Golnazsadat Zargarian
Data scientist

Golnazsadat.zargarian@altran.it