MEETUP
RADIX

Tema: Machine Learning
Palestrante: Raul Sena Ferreira

radix

BIO

Data Scientist at Radix

Post-graduate Lecturer at INFNET

MSc., Engineering Systems and Computing (Machine Learning) - UFRJ

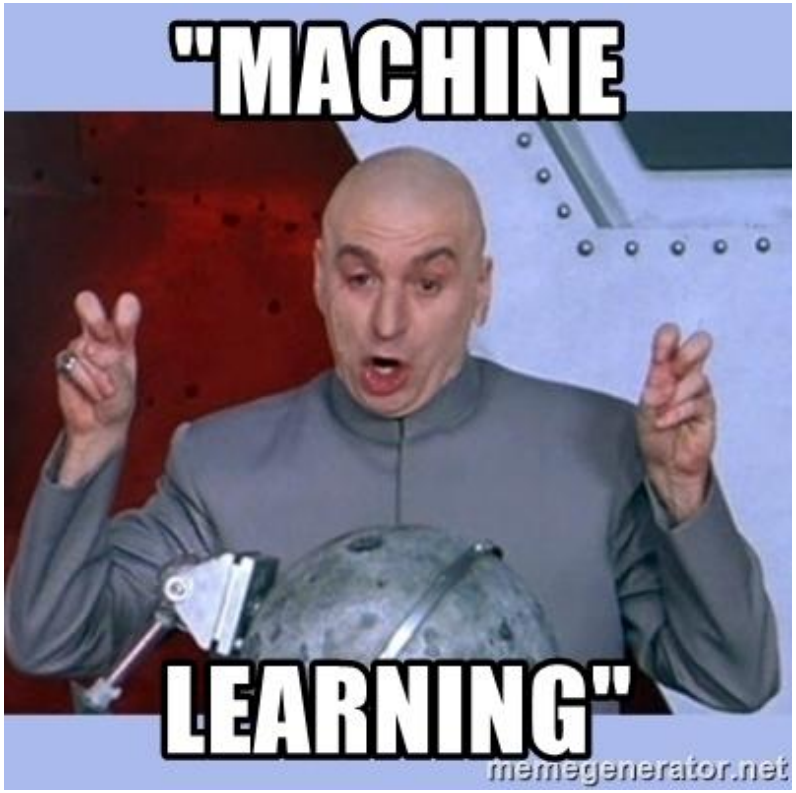
BSc., Computer Science - UFRRJ

http://www.raulferreira.com.br/

https://github.com/raulsenaferreira/Talks-and-Presentations/

radix

What is Machine Learning ?



A subfield from artificial intelligence

It uses statistical techniques for making computers to learn from data and act like humans do in an autonomous way

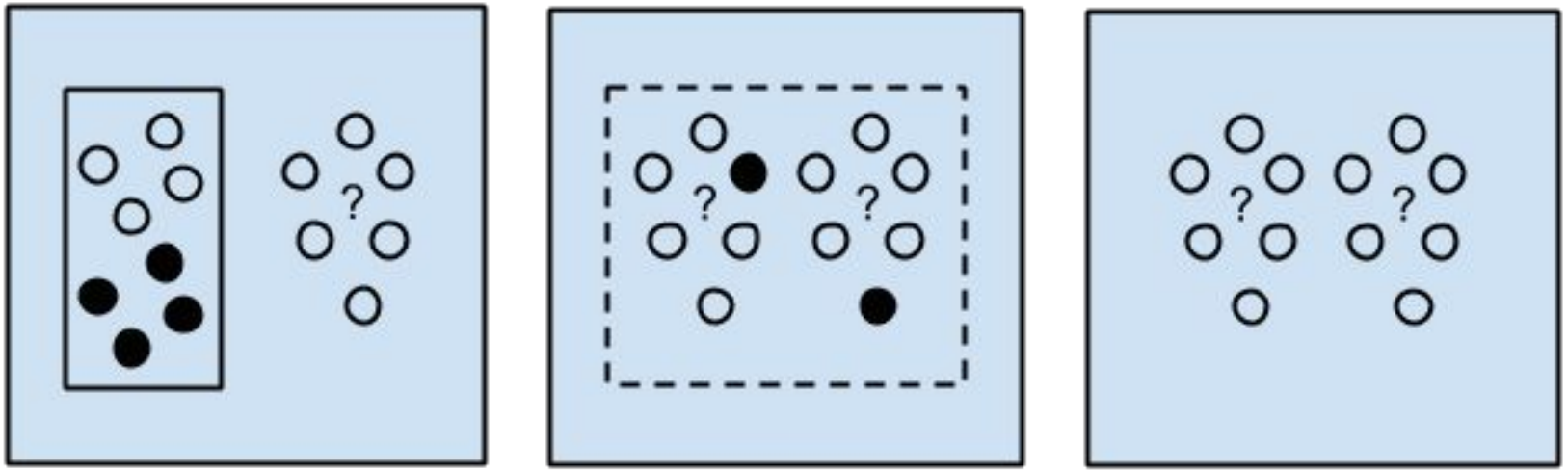
- Statistics
- Data Mining techniques
- Optimization and heuristics

radix

Most common techniques

- Classification
- Regression
- Clustering

Three main ways to tackle a problem



radix

Advantages

Automation of tasks

Fast Processing and Real-Time Predictions

Good feature representations for a given task

Applications:

- Text categorization
- fraud detection
- optical character recognition
- market segmentation
- natural-language processing
- machine vision ...

Is the machine intelligent ?

radix

Drawbacks

ML needs a lot of training data for future prediction

It is not a guarantee that ML will always work in every case

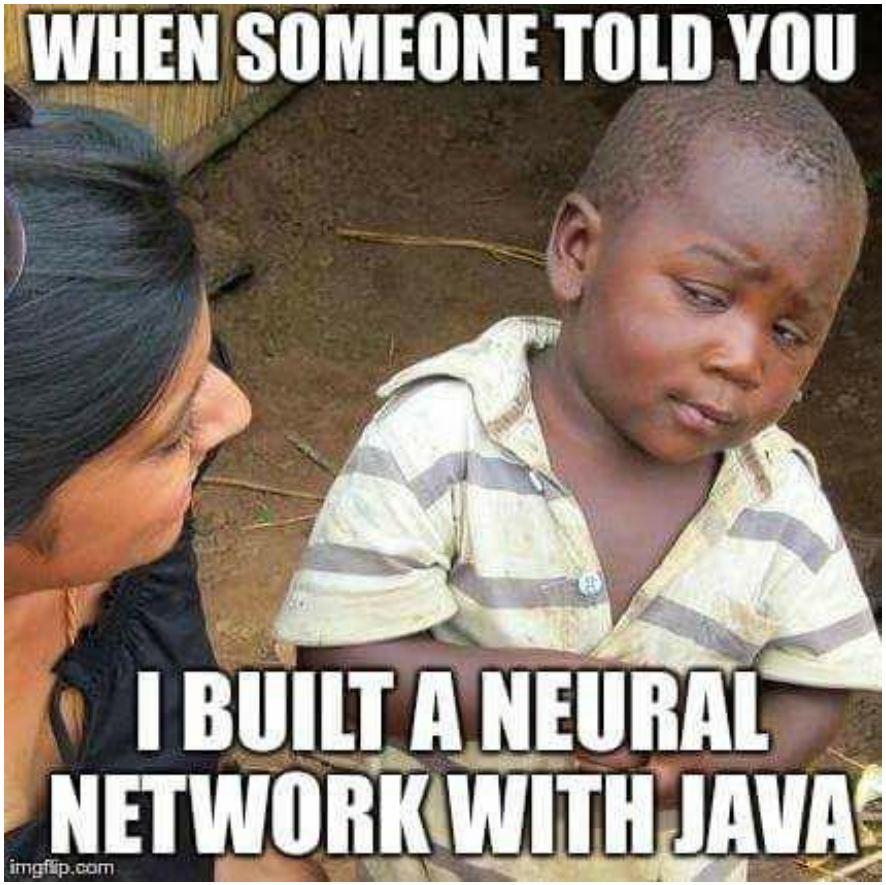
Statistical approach are more reliable in some scenarios

Low interpretability (domain knowledge is still needed)

A lot of pre-processing work has to be done

Business Intelligence are more suitable for certain tasks

Let’s talk about what we use for ML



Learn Python !

- Scikit learn
- Numpy
- Pandas
- Scipy
- Matplotlib
- Pre-processing tools

Python for the entire pipeline

- Flask (API)
- Django (System)

Checklist for a machine learning project

- Classification or regression ?
- Dataset: balanced or unbalanced ?
- Supervised or unsupervised ?
- High dimensional dataset ?
- Is it an online or an offline domain ?
- Big data problem or a small data one ?
- Has the dataset too much noise ?

And other business questions ...

And after all of that ? What to do first ?

KEEP
CALM
AND
TELL ME BABY
WHATS YOUR STORY?

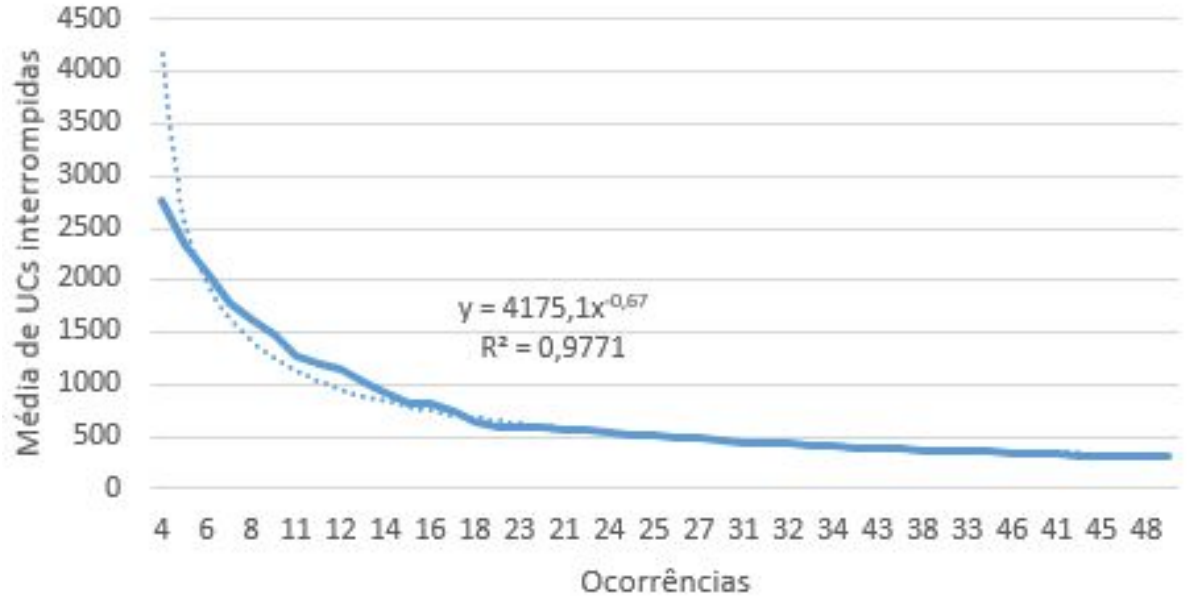
radix

Data storytelling

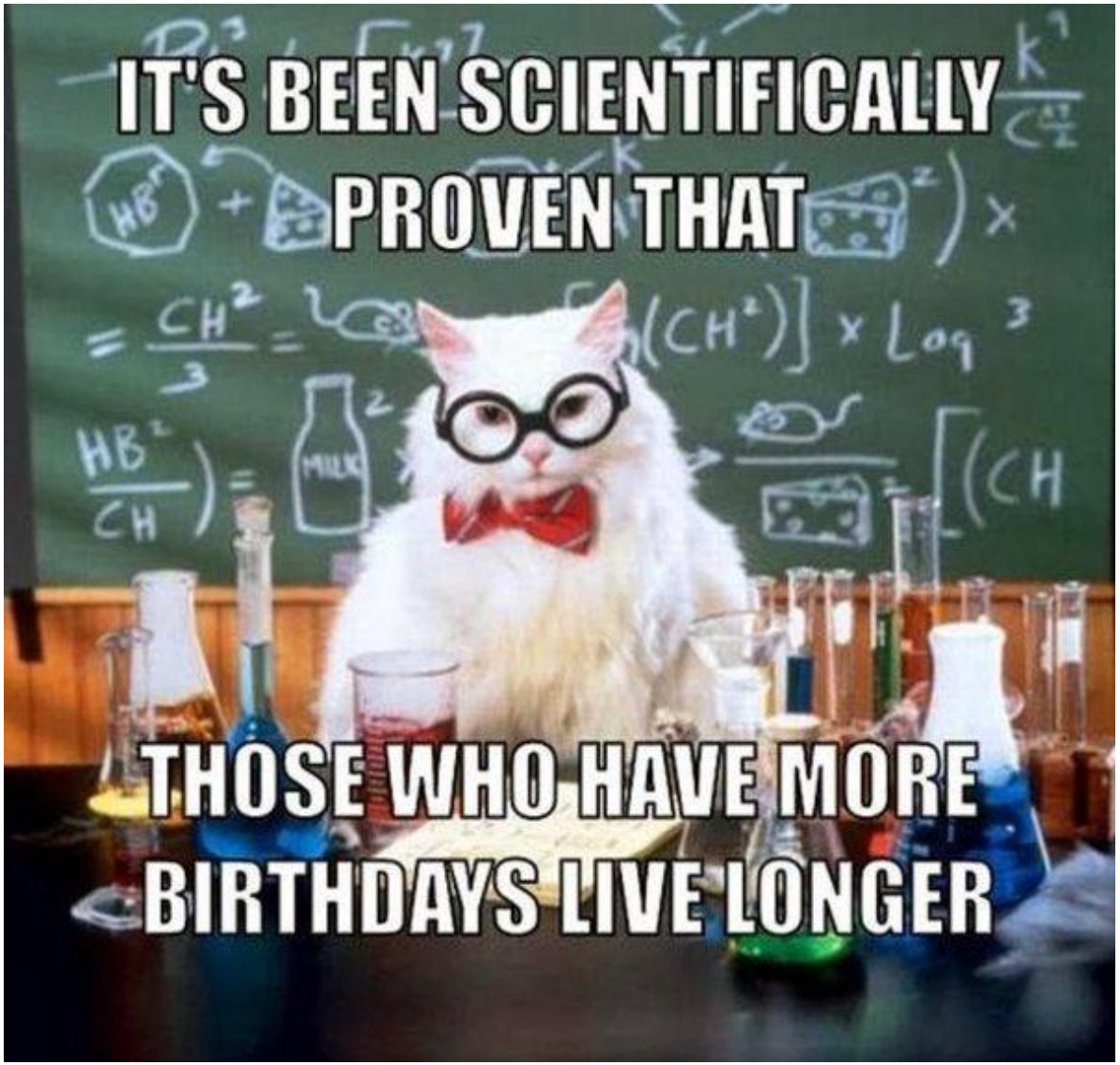
Jupyter for exploratory analysis

- Cells avoid reloading content
- Cells allow ease of explanation and analysis

How is distributed your data ?



WOW! You discovered something in your data !



Pre-processing

Categorical features to numerical ones

- Label encoder; one-hot encoder

Dealing with missing data

- Representing “no information”

Coding dictionaries and/or keywords from dataset

- Prepare data for a better algorithm processing

Feature engineering

Correlation between attributes

- Values near from -1 or 1 (discard one of the features)

Feature importance

- Tree-based algorithms give the features importance

Information extraction

- An attribute may have other aggregated attributes
- Hidden attributes through PCA or SVD for latent variables

Feature scaling

- Your attributes have huge difference between each other

radix

Simple models are preferred

Picking a set of algorithms

After answering all of the previous questions

- Baseline algorithm (traditional and easy to explain)
- ~5 algorithms to test against the baseline

Tree-based algorithms are robust and widely used

- Random forest; XGBoost; ADA Boost …

Traditional algorithms are strong in many types of domains

- Neural networks
- K-NN
- Linear Regression

Tuning your model

Optimization methods for tuning the hyper-parameters

- Random search for wide range of values
- Grid search for fine tuning within a small range

Who is the best? Your fancy algorithm or the simple baseline ?

Do you know about null hypothesis ? Statistical significance ?

radix

Evaluating ML models

Classification

- Accuracy
- MCC
- F1

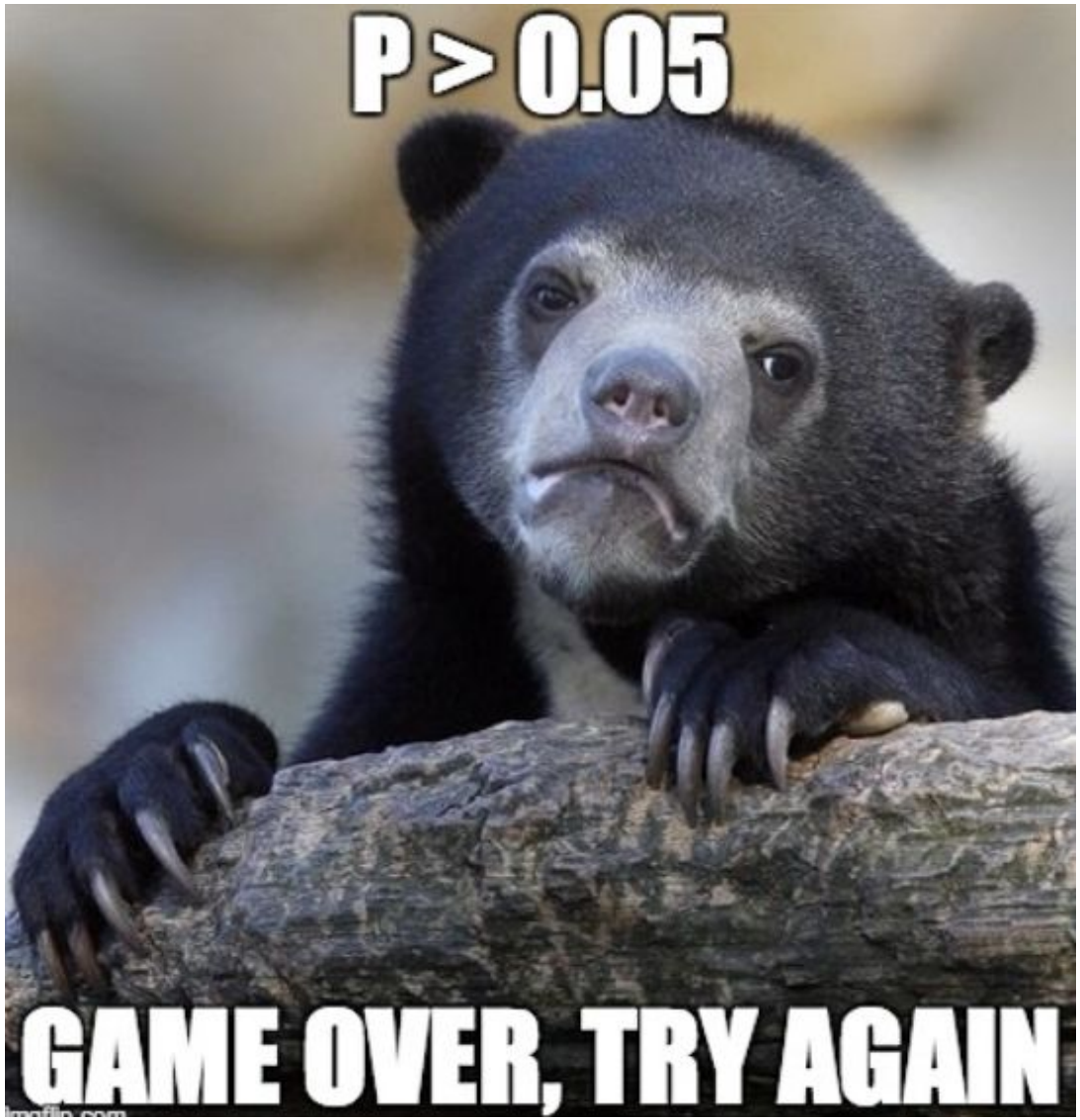
Regression

- MAE
- RMSE
- R^2 (chi squared)

Hypothesis tests for validating real improvements !

radix

Is there any statistical difference ?



Don’t give up

Repeat all steps until you get better results than baseline

Exhausted all temptations?

- Choose the baseline

Do we have a tie between the models?

- Umn... err... choose the baseline :)

Don’t have a pet algorithm !

Convincing the customers

Documentation

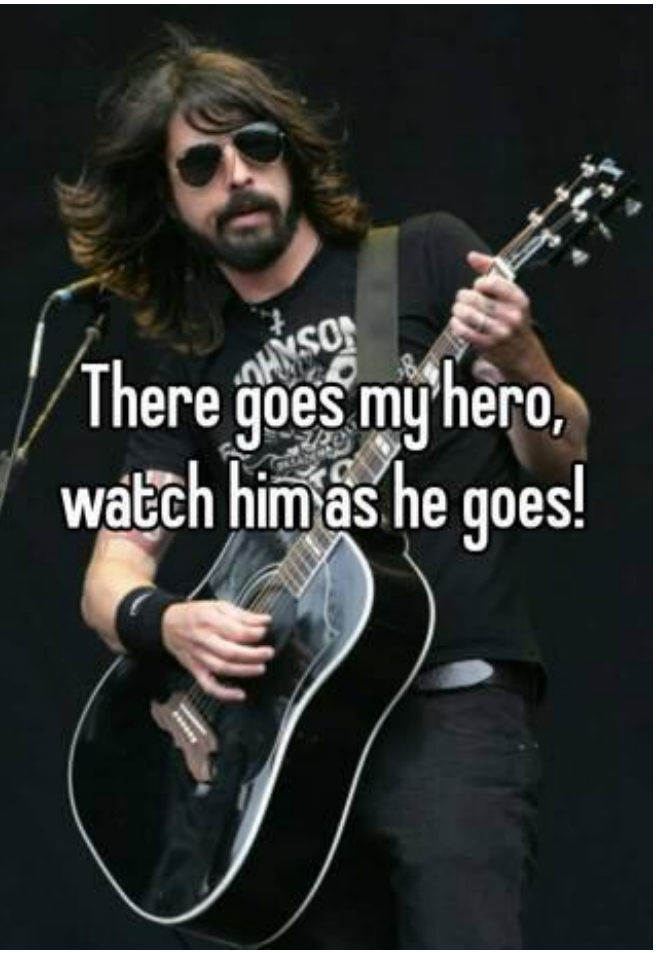
- Experiments Methodology (for reproducibility purpose)
- Features list (and what you did to remove some of them)
- List of installed libraries (pip install requirements ?)

Plots

- Scores (and the meaning of the metrics)
- Confusion matrix / AUC graph
- Features importance
- New quantitative results compared against old numbers

Everything fine ? Time to move our code to production env.

Before production



Saving an object in a binary format

- **pickle.dump(model,open("DI_Regressor.pickle.dat", "wb"))**

Loading a saved object

- **pickle.load(open("DI_Regressor.pickle.dat", "rb"))**

Log for everything with **logging** library

- **logger = logging.getLogger('regression')**
  logger.error('Error: ' + str(error))

Shipping to production

It’s preferable all machine learning code in one file

- Simplicity rather than complexity
- Vectors and matrices rather than objects

Flask can be used as web API (not mandatory)

- A traditional strategy for systems can be used

Server requirements documentation

- Along with memory and processing test stress analysis

Save all predictions in a table for posterior comparison

radix

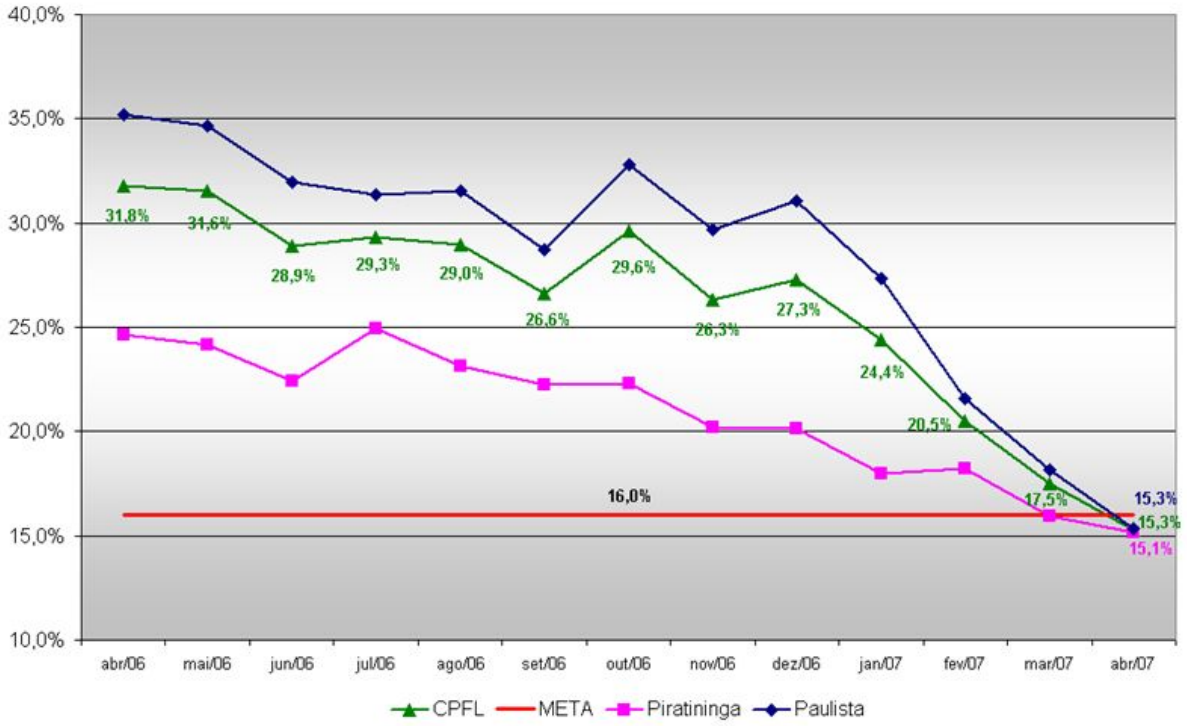
What are we doing with ML right now ?

Maintenance teams are dispatched to fix problems that is not concerned to the CPFL company

R$136.00 per call

- ~14 Millions of costs

How to analyze the customer calls and classify who needs to be attended first



First results

The machine detects 79% of undesirable dispatches

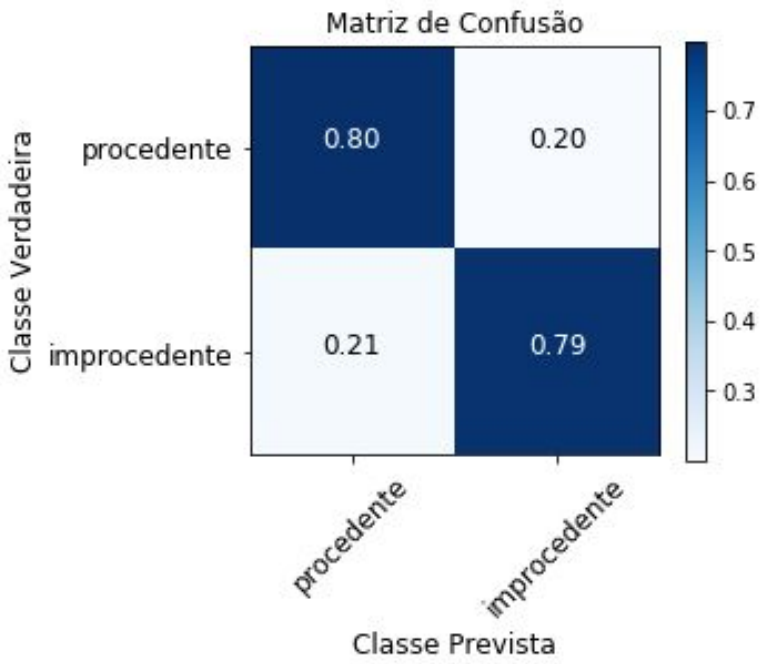
- Actual human accuracy is 37%

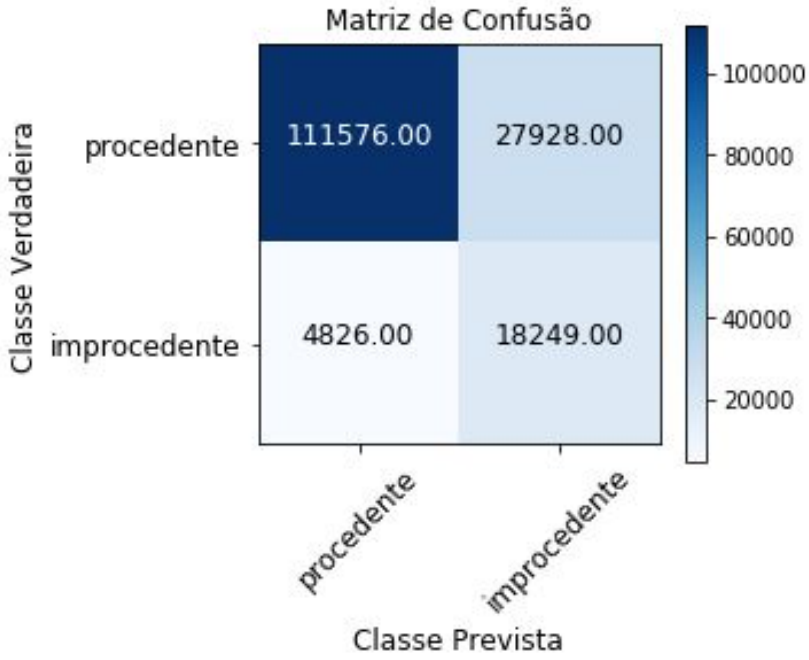
Machine detects 80% of all correct dispatches (500K events)

- Human detects 100% from a portion

A human can take up to 5 minutes to take a one decision

The machine takes 13 seconds to make 50,000 decisions





Under development

Virtual analyzers

- How to predict the fluids indices ?

Speed for laboratory analyzes and less efforts for professionals

Predictive maintenance

- How to analyze temporal signals regarding the valves and predict when will occur a failure ?
- How to predict the amount of time until a new failure ?

Reducing maintenance costs and time

Increased demand for ML projects

Currently staff

- 4 machine learning researchers
- 4 aspiring machine learning researchers

4 ongoing projects and 3 other in prospection

- Huge potential of several future projects
- Actual projects with potential of extended work

Broad range of applications and lack of intelligent solutions

- Brazilian market & industry are discovering ML potentials
- After big data, ML is now the new buzzword

radix

Hot topics in machine learning

Large scale Machine Learning

Deep Learning (Explainability)

Reinforcement Learning

Computer Vision

Internet of Things

radix

I loved everything ! How can I start ?

Online courses

- https://www.coursera.org/learn/machine-learning
- https://www.edx.org/learn/machine-learning
- https://www.udemy.com/topic/machine-learning/

Books

- Elements of statistical learning
- Machine learning: A probabilistic perspective
- Python: Data structures and algorithms

radix

DÚVIDAS?

CONTATO

Raul Sena Ferreira
raul.ferreira@radixeng.com.br
raulsf@cos.ufrj.br