AirQuality

Nicolas Bohorquez
MADAS - Smart Cities Lab
Prof. Claudio Rossi

AirQuality **Prediction**

From messi(y) to results

At the beginning of 2014, Telecom Italia launched the first edition of the Big Data Challenge, a contest designed to stimulate the creation and development of innovative technological ideas in the Big Data field.

The AirQuality Dataset describes the pollution type and intensity of Milan city using various types of sensors located within the city limits. Also datasets about weather conditions and road transportation traffic were released.

Original data came as several Comma Separated Values (CSV) files. It comprises 61 days of hourly taken observations.

AirQuality Prediction

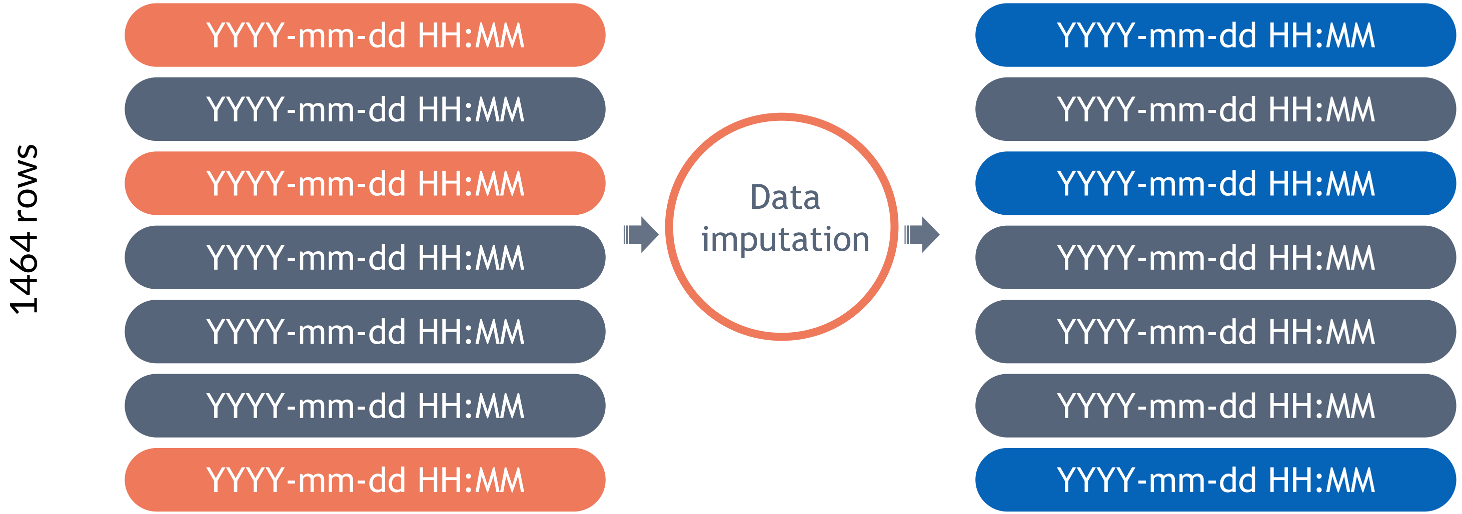
From messi(y) to results

Using data provided with some basic Machine Learning Techniques we built a model to classify the hourly observations into one of the 7 categories (1 = Optimus to 7 = Very unhealthy) defined by the “Agenzia Regionale Prevenzione e Ambiente” for the Piemonte, Italy.

We compare three distinct common classification algorithms (**LogisticRegression**, **RandomForest**, **GradientBoostingClassifier**) with different groups of features, finally we used an automatic technique (**Recursive feature elimination with cross-validation**) to improve the manually selected set of features.

Messy **Data**

Using value assignment* and interpolation to enrich the data



(*strong assumption)

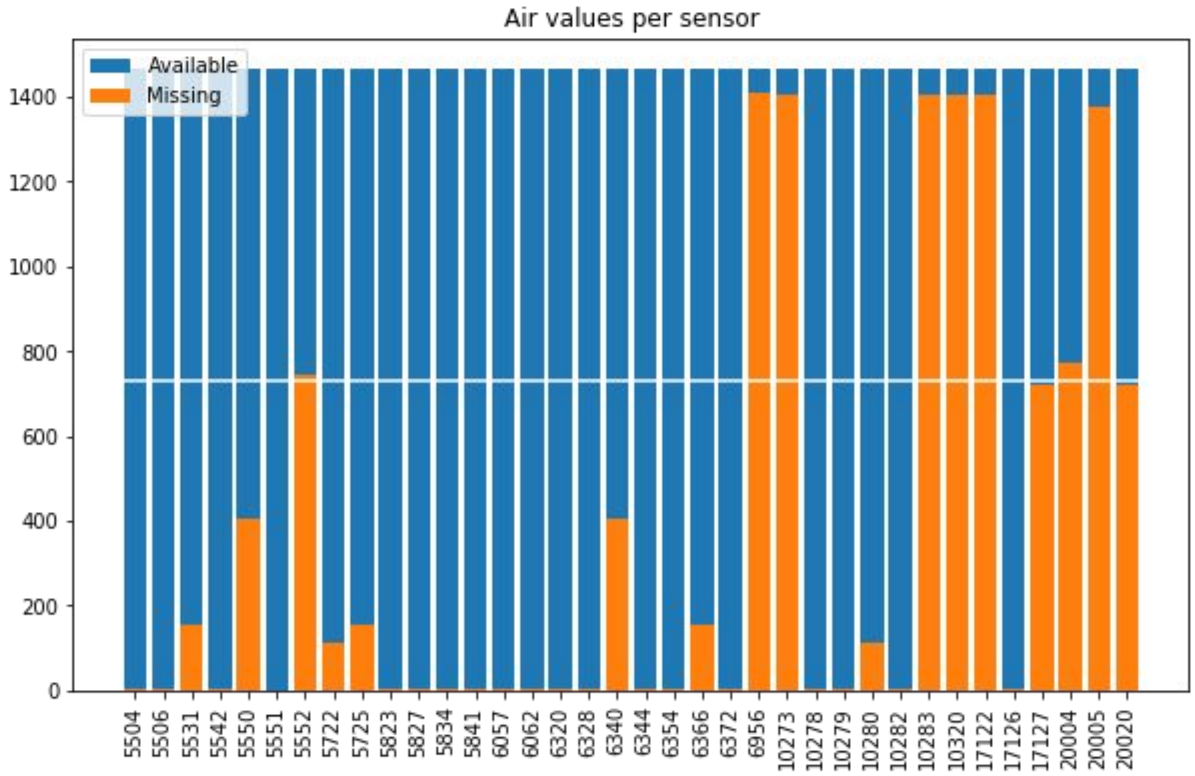
Messy Data

Using value assignment* and interpolation to enrich the data

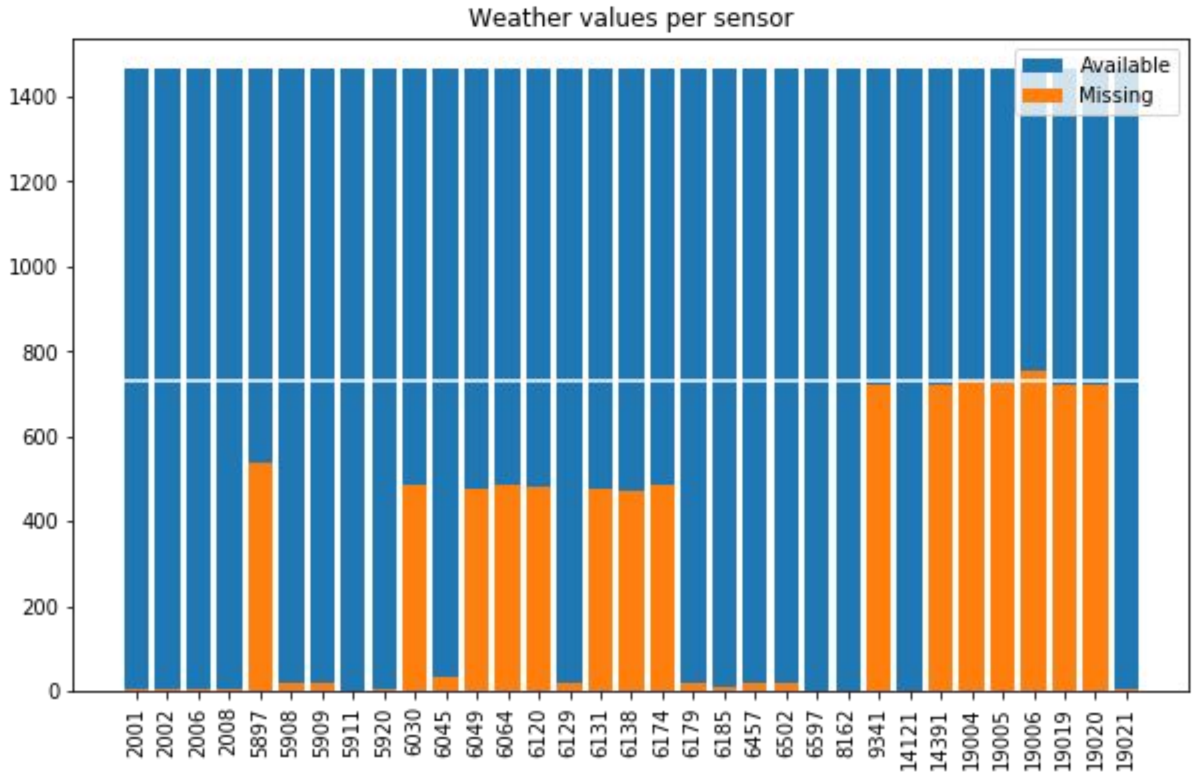
Every column

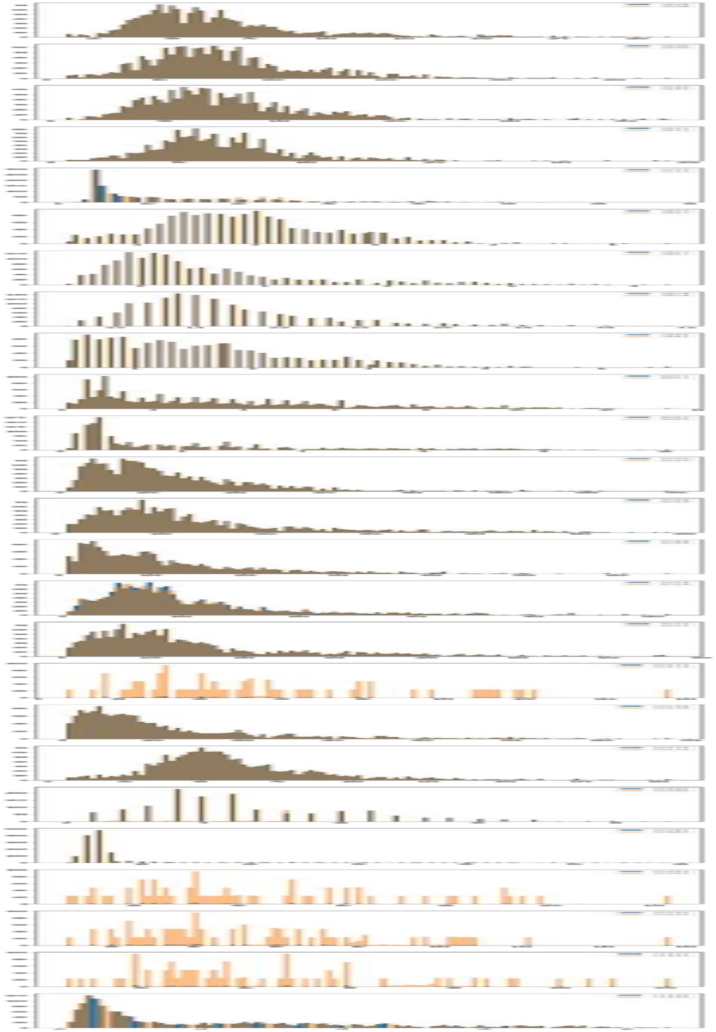
- Missing hours > 1464 / 2 => ignore column
  - Missing hours in a day == 23 => assign the same value for all hours of the day
    - Missing hours <= 10% => interpolate (spline, forward-backward, max 3 values)
    - Ignore column

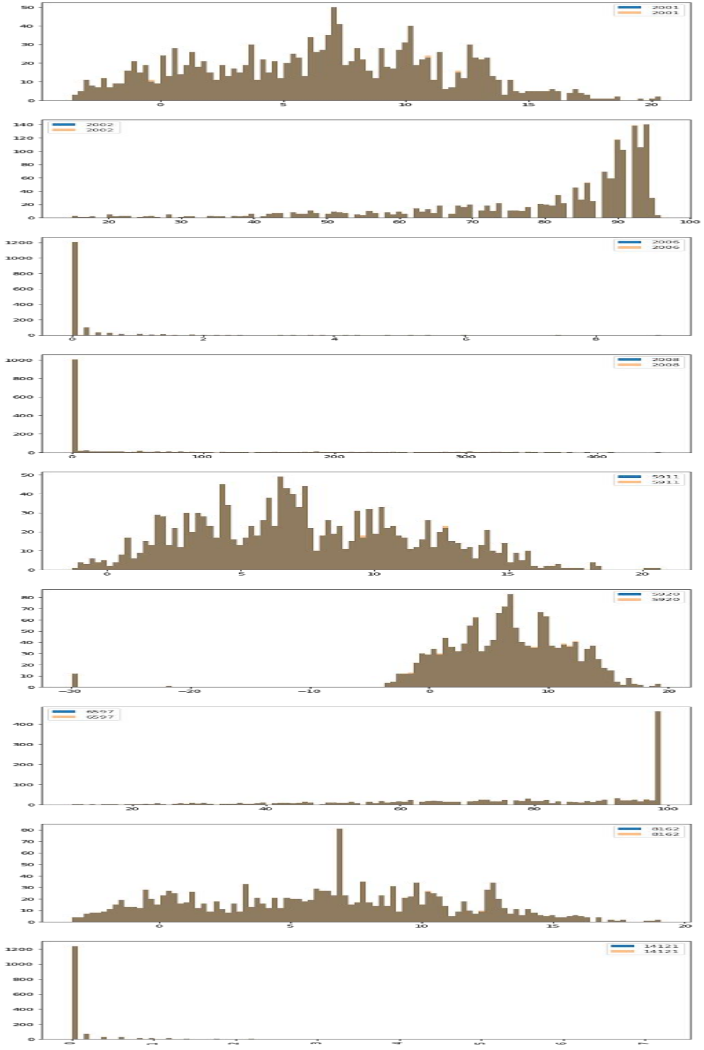
(*strong assumption)



Air values per sensor
Available
Missing
0
200
400
600
800
1000
1200
1400
5504
5506
5531
5542
5550
5551
5552
5722
5725
5823
5827
5834
5841
6057
6062
6320
6328
6340
6344
6354
6366
6372
6956
10273
10278
10279
10280
10282
10283
10320
17122
17126
17127
20004
20005
20020


Weather values per sensor
Available
Missing
0
200
400
600
800
1000
1200
1400
2001
2002
2006
2008
5897
5908
5909
5911
5920
6030
6045
6049
6064
6120
6129
6131
6138
6174
6179
6185
6457
6502
6597
8162
9341
14121
14391
19004
19005
19006
19019
19020
19021




2001
2001
2002
2002
2006
2006
2008
2008
5911
5911
5920
5920
6597
6597
8162
8162
14121
14121

Feature Construction

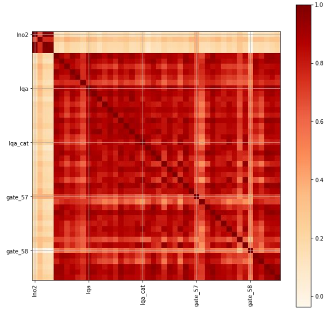
Iterative approach.

CLASSIFICATION

This includes the selected features from automatic / manual selection, accuracy vs calculated IQA

Weather Conditions

Vehicle Characteristics



Traffic by **Gate**

Features available

The Algorithms

Logistic Regression.

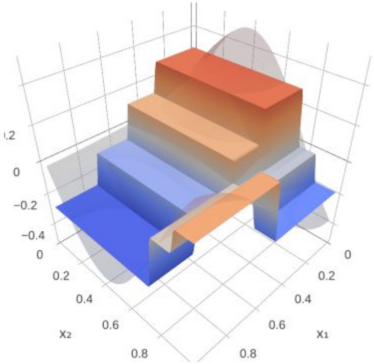
Used as benchmark, without special tuning to validate the assumptions of an OLS,

Random Forest.

Fast classification, a way to tackle the overfitting problem, no special configuration.

Gradient Boost.

Fast classification, a way to tackle the overfitting problem, no special configuration.



Sum predictions of an ensemble of trees.

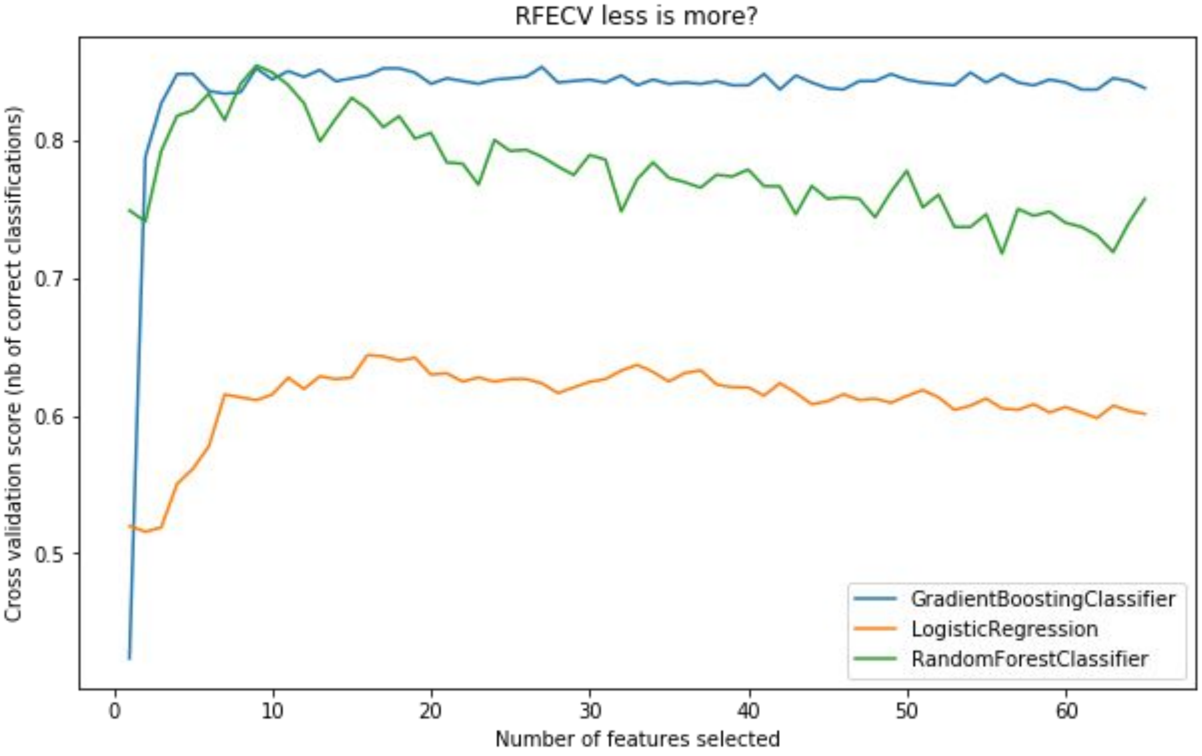
$$D(\mathbf{x}) = d_{tree1}(\mathbf{x}) + d_{tree2}(\mathbf{x}) + d_{tree3}(\mathbf{x})$$

Next tree should complement and and minimize the training error of the ensemble.

$$D(\mathbf{x}) + d_{tree4}(\mathbf{x}) = f(x)$$

Automatic Feature selection

Improving our intuitions.

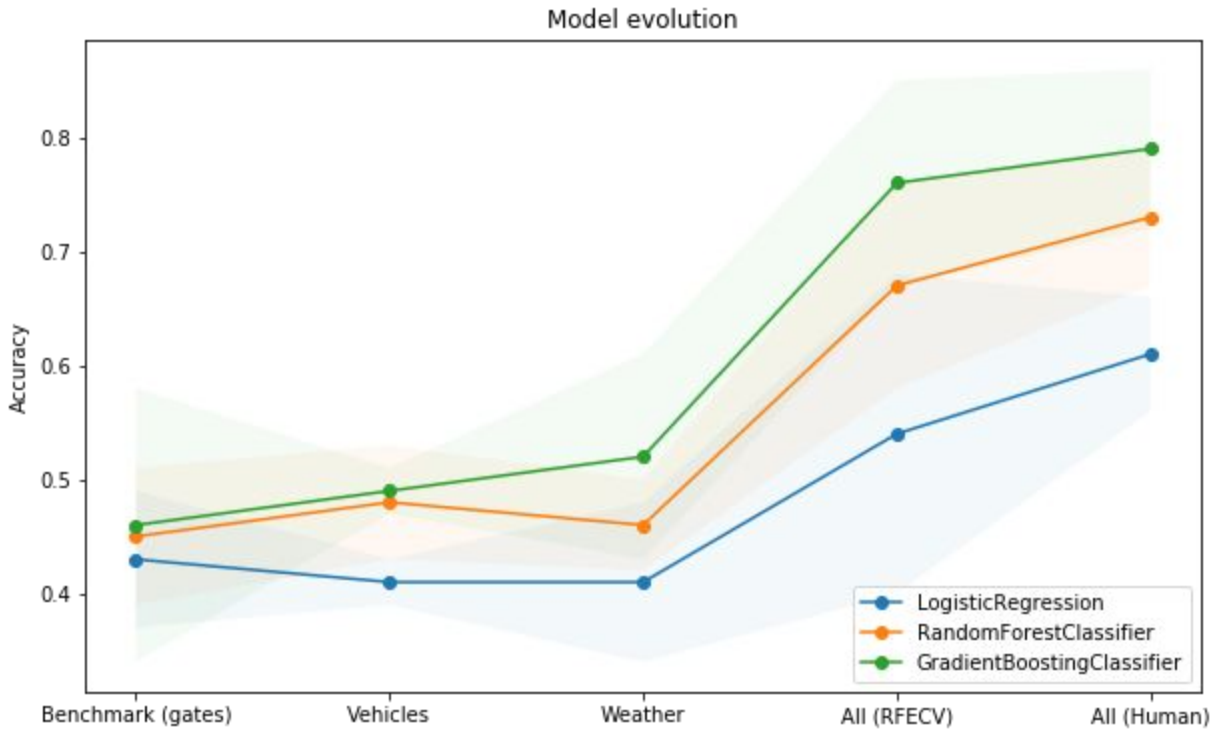


"Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), recursive feature elimination (**RFE**) is to select features by recursively considering smaller and smaller sets of features."

scikit-learn v0.19.1 design for machine learning software: experiences from the scikit-learn project

Results Comparison

Avg Score classifying the AIQ (Area shows +/- std).



Lessons **Learned**

Personal | General

Statistics
Need to improve radically the understanding of the basis

Algorithms
Statistics + Creativity + Coding

Curiosity
Temporal networks (directed graphs) that combines wind direction, traffic and vehicles characteristics?

- Hard to find a "***correct***" imputation
- Data preparation takes **75%** time
- Feature selection is cool but difficult
- There are lies, damn lies and statistics

Thanks for the Project