Adrien Savary

Failure risk prediction on pipeline network

1.

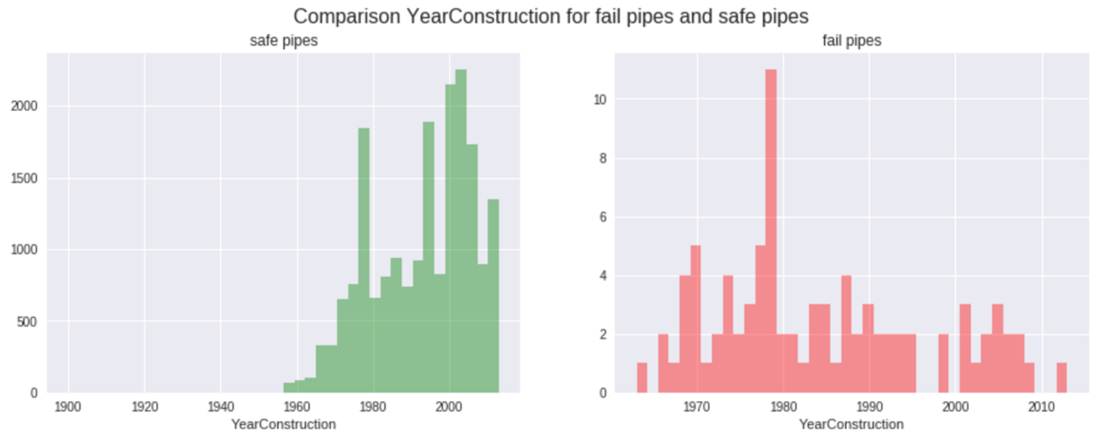
Exploratory Analysis

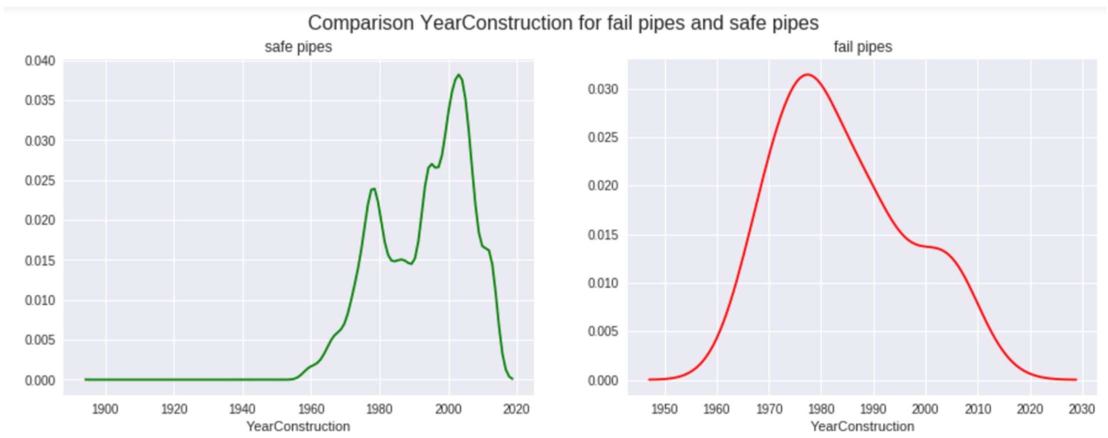
First Look at the Data

Data set

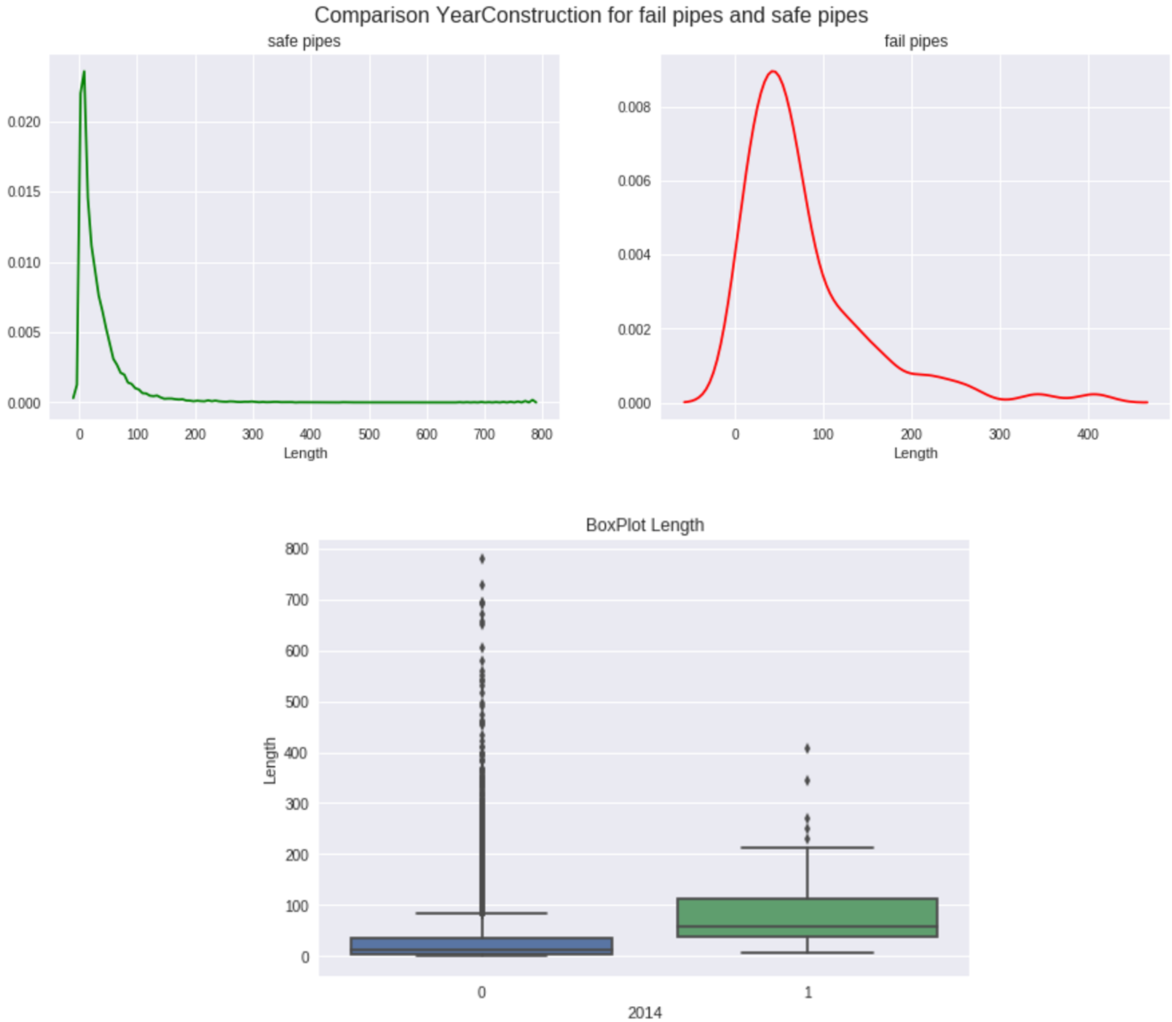
- 19428 pipes
- 7 features : low dimensionality
- We know pipes that will fail in 2014 and 2015
- Very imbalaced
- 2014 (0.27% failed)
- 2015 (0.19% failed)
- Predict probabilities of failure for 2014 and 2015

YearConstruction





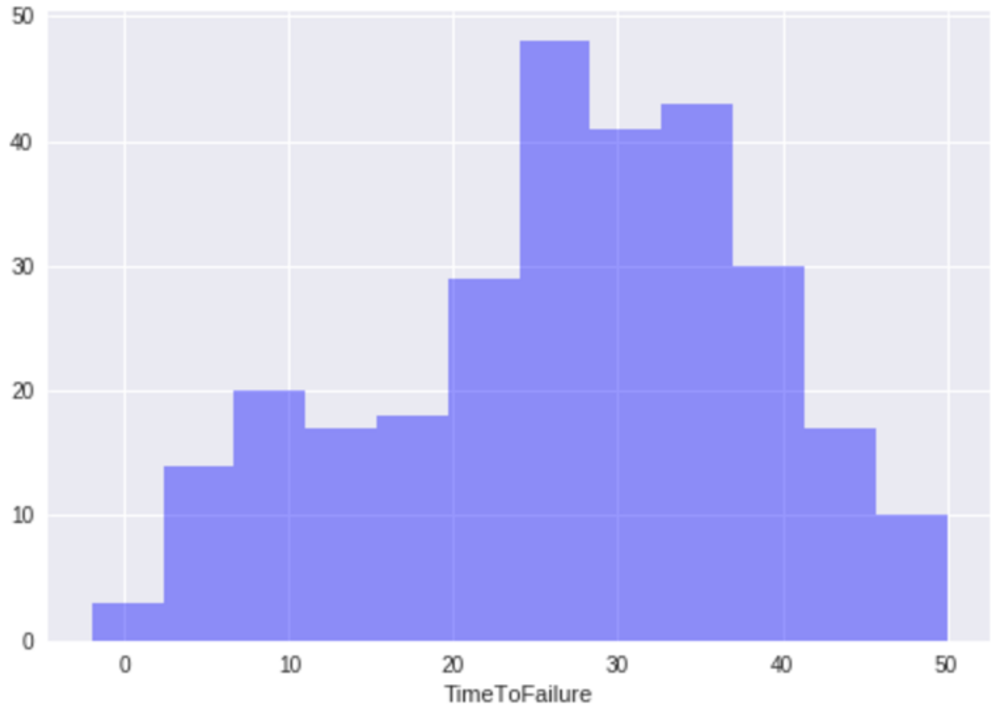
Length



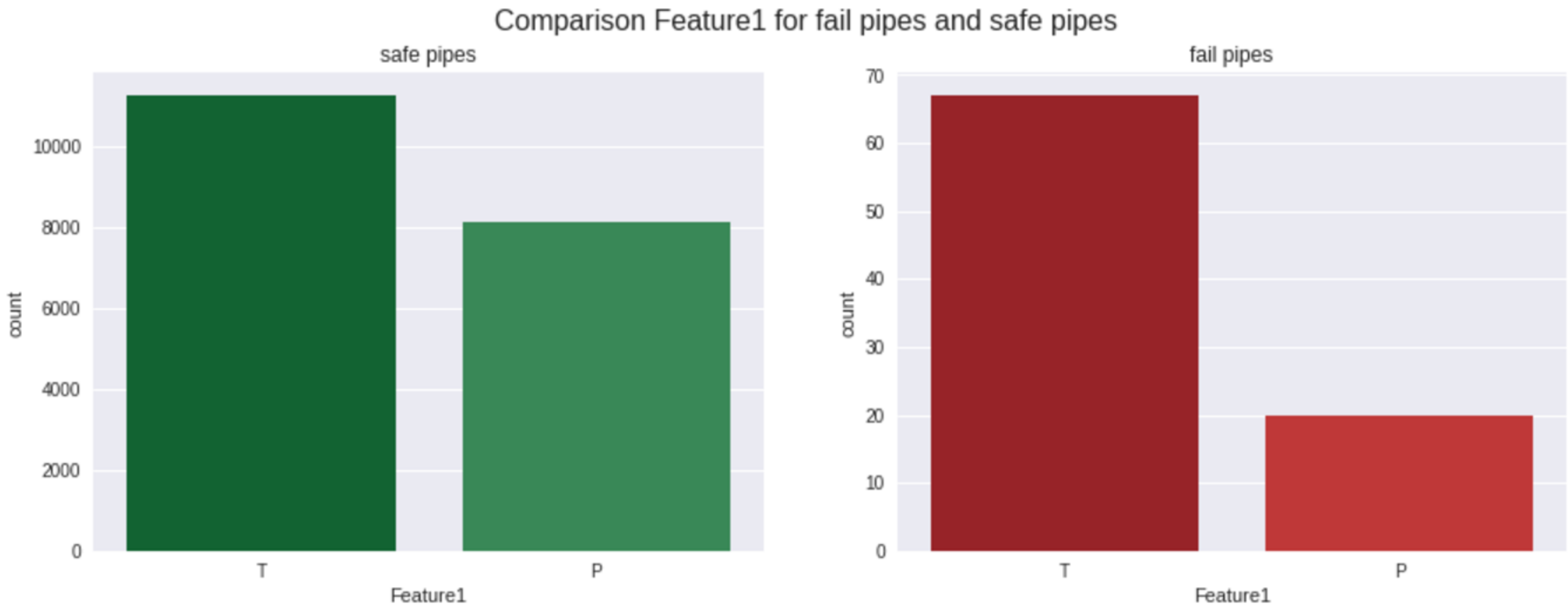
Comparison YearConstruction for fail pipes and safe pipes
safe pipes
fail pipes
0.020
0.015
0.010
0.005
0.000
0 100 200 300 400 500 600 700 800
Length
0.008
0.006
0.004
0.002
0.000
0 100 200 300 400
Length
BoxPlot Length
800
700
600
500
400
300
200
100
0
Length
0
1
2014

TimeToFailure

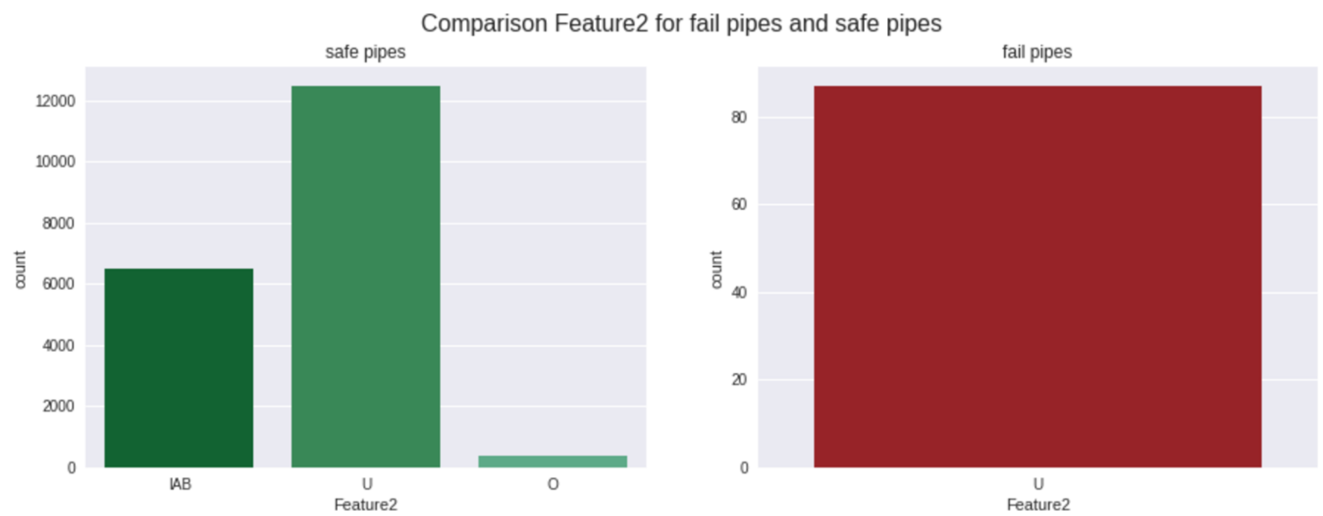
- New Feature to get more insights
- TimeLastFailureObserved – YearConstruction
- Censored data



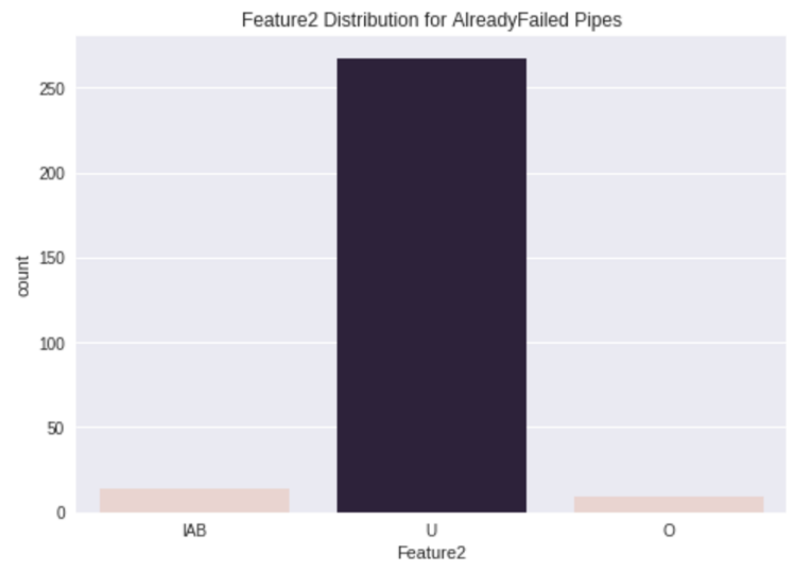
Feature1


Comparison Feature1 for fail pipes and safe pipes
safe pipes
count
10000
8000
6000
4000
2000
0
T
P
Feature1
fail pipes
count
70
60
50
40
30
20
10
0
T
P
Feature1

Feature2

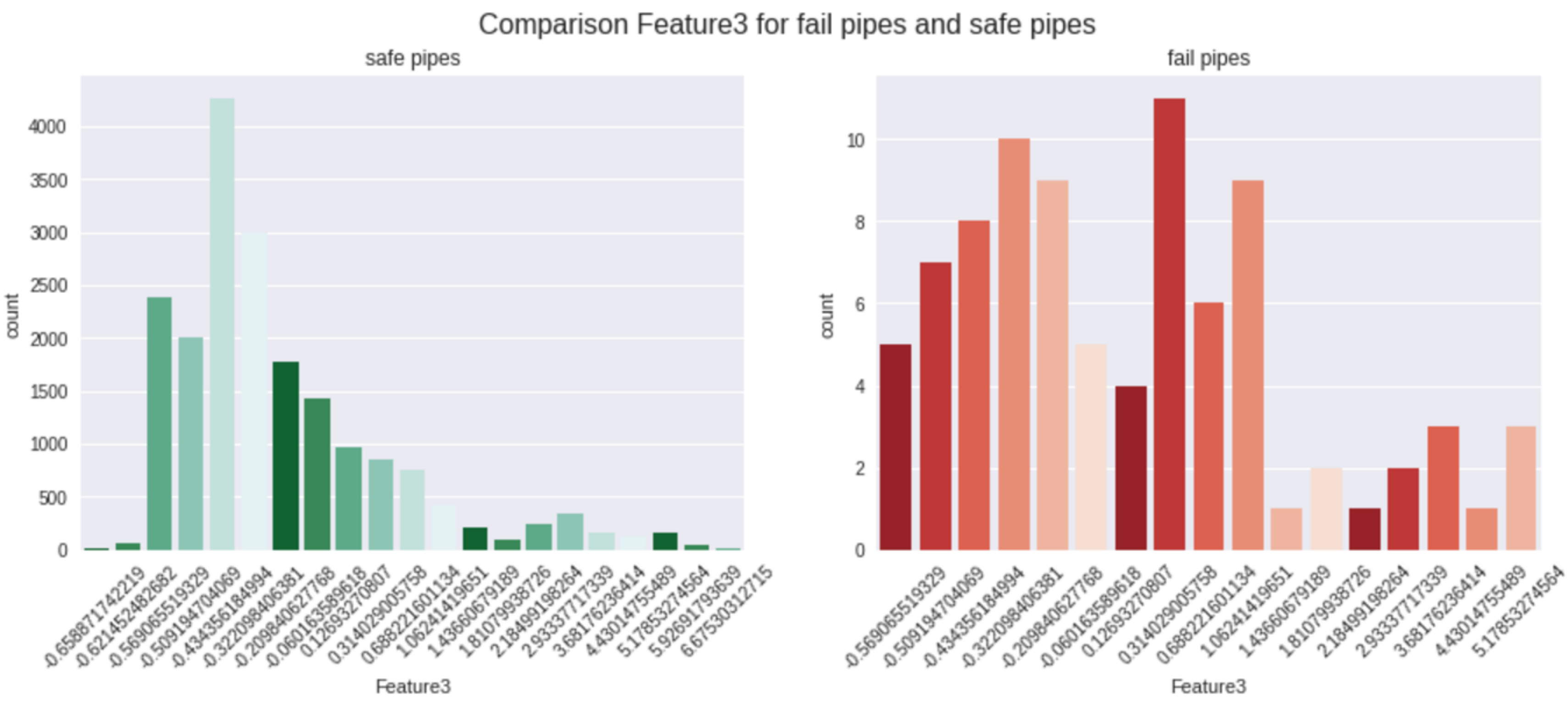


Comparison Feature2 for fail pipes and safe pipes
safe pipes
fail pipes
count
12000
10000
8000
6000
4000
2000
0
IAB
U
O
Feature2
80
60
40
20
0
U
Feature2



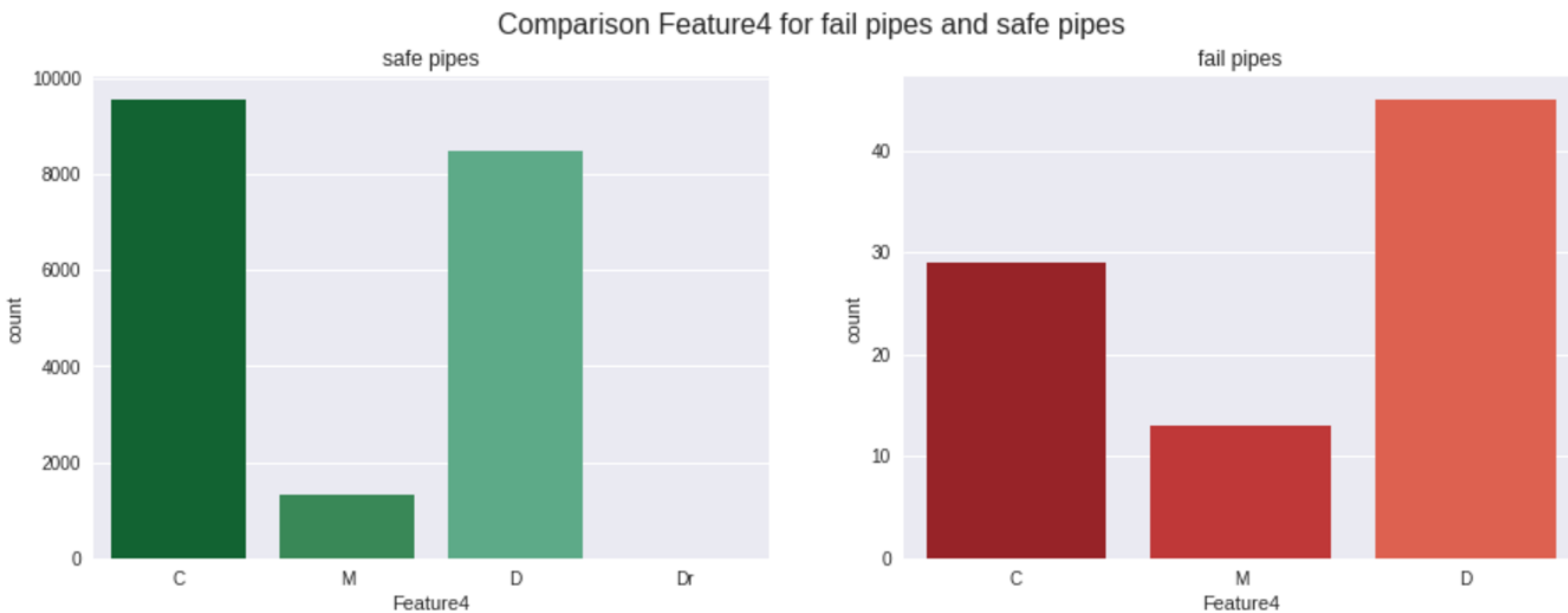
Feature2 Distribution for AlreadyFailed Pipes
count
250
200
150
100
50
0
IAB
U
O
Feature2

Feature3



Comparison Feature3 for fail pipes and safe pipes
safe pipes
count
4000
3500
3000
2500
2000
1500
1000
500
0
-0.658871742219
-0.621452482682
-0.569065519329
-0.509194704069
-0.434356184994
-0.322098406381
-0.209840627768
-0.060163589618
0.12693270807
0.314029005758
0.688221601134
1.06241419651
1.43660679189
1.81079938726
2.18499198264
2.93337717339
3.68176236414
4.43014755489
5.17853274564
5.92691793639
6.67530312715
Feature3
fail pipes
count
10
8
6
4
2
0
-0.569065519329
-0.509194704069
-0.434356184994
-0.322098406381
-0.209840627768
-0.060163589618
0.12693270807
0.314029005758
0.688221601134
1.06241419651
1.43660679189
1.81079938726
2.18499198264
2.93337717339
3.68176236414
4.43014755489
5.17853274564
Feature3

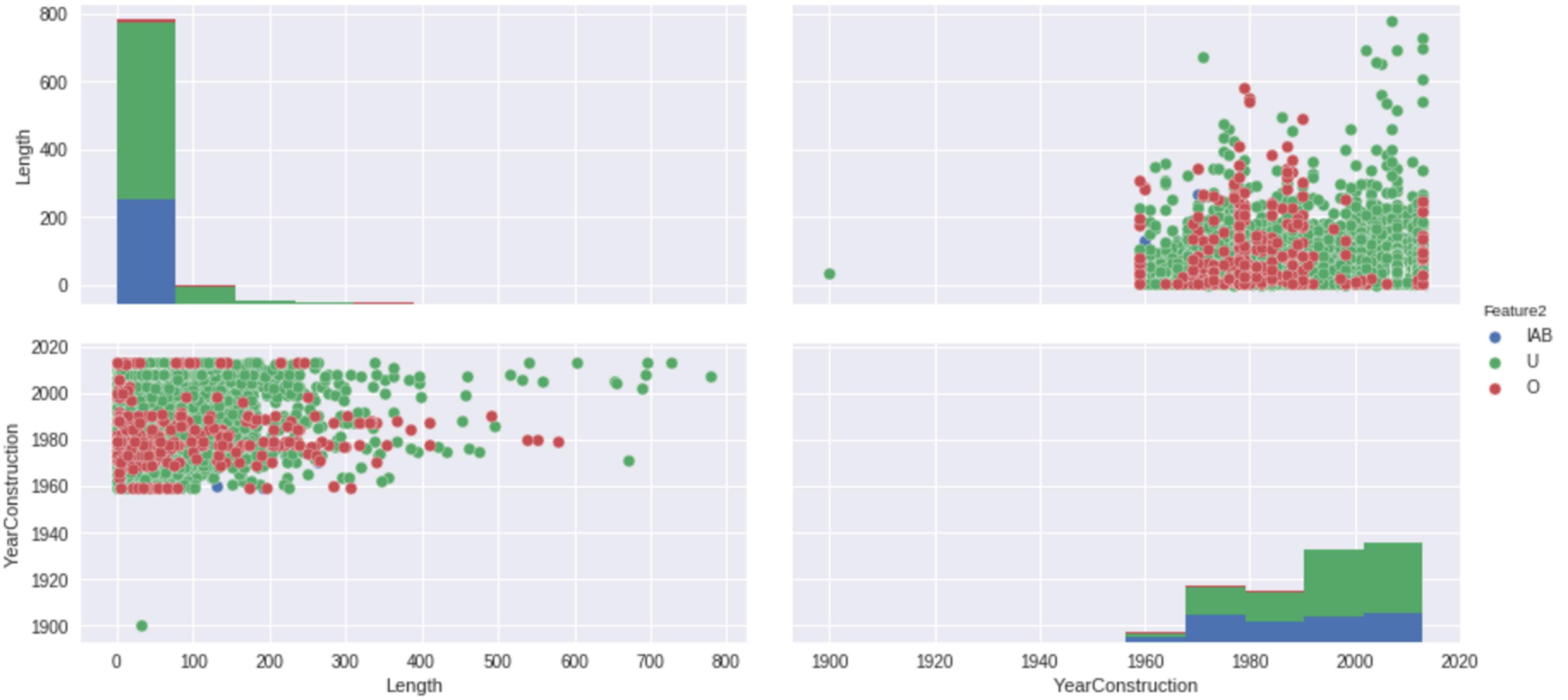
Feature4



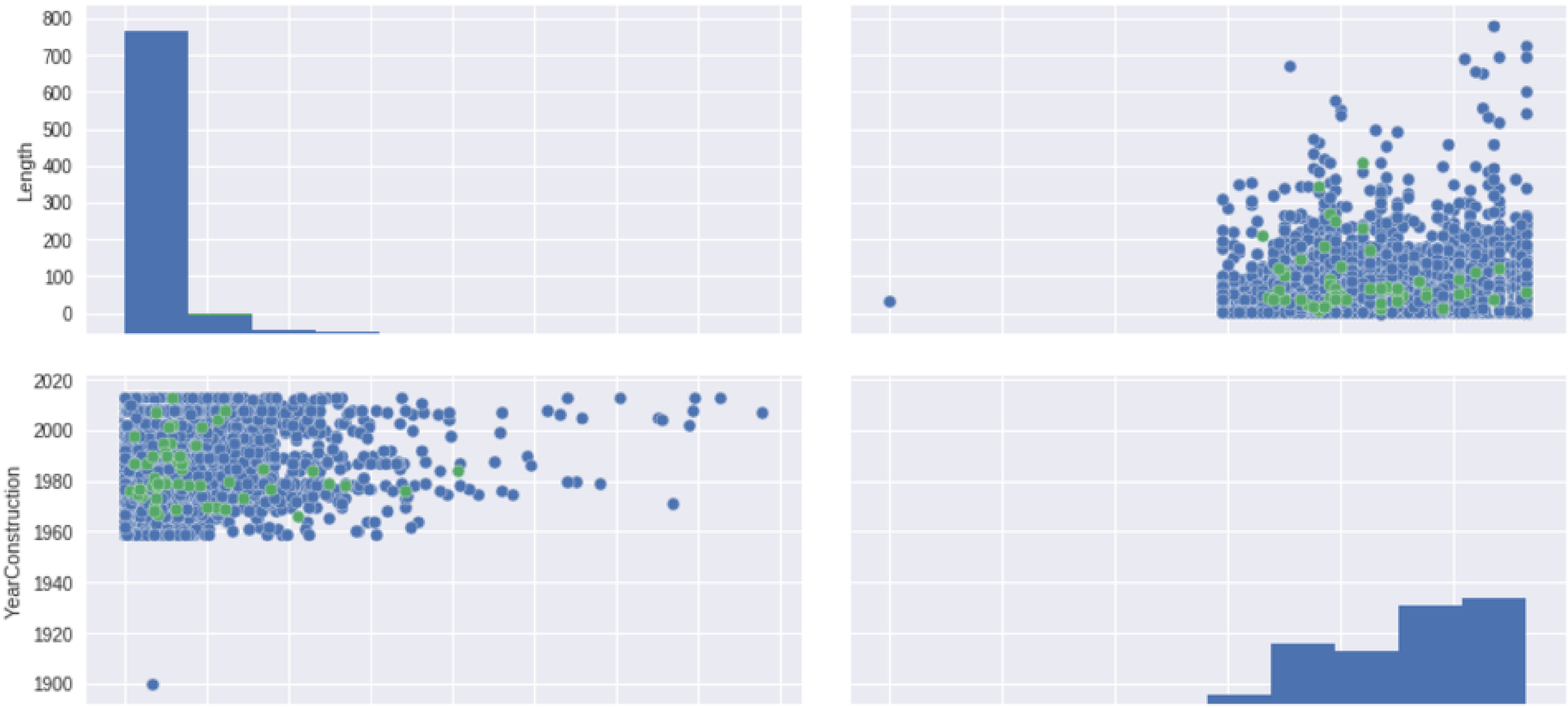
Comparison Feature4 for fail pipes and safe pipes
safe pipes
10000
8000
6000
4000
2000
0
count
C
M
D
Dr
Feature4
fail pipes
40
30
20
10
0
count
C
M
D
Feature4

Pairwise relationship

- Feature O : generally older
- Feature U : generally longer



Length/YearConstruction by failures



▷ Green : fail pipes for 2014

2.

Prediction methods

Supervised Learning for imbalanced data

Challenge Metric: ROC-AUC

- Area under the ROC curve
- The closer to 1 the better
- If we pick a random positive and a random negative, the ROC-AUC gives the probability that a classifier assigns a higher score to the positive example

Varun Chandola in Anomaly detection a survey

“

Normal points occur in dense regions while anomalies occur in sparse regions.

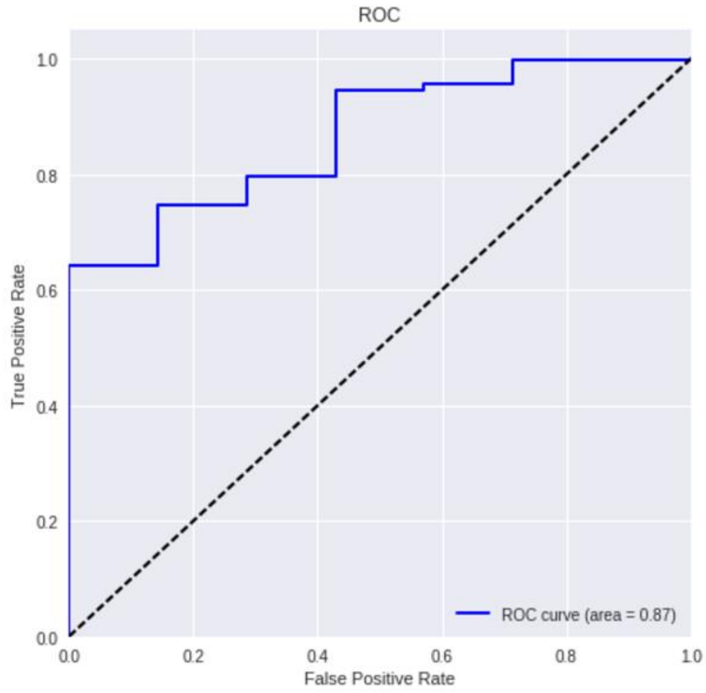
Normal point is close to its neighbours and anomaly is far from its neighbours.

Simple Anomaly detection (semi-supervised)

- Density estimation using only continuous features
- Fit a gaussian mixture model to safe pipes data
- Prediction with very low probability would be an outlier
- Not robust, depends highly on the initialization
- Around 80% ROC-AUC

Simple Logistic Regression

- ▷ 87% cross-validation
- ▷ 85% Test



Undersampling

- ▷ **General idea**
  - Train a classifier on a smaller sample of the data
  - The sample is balanced
  - Force the classifier to put more weight on outliers
  - Problem : You lose data

- ▷ **What we tried**
  - Randomly select balanced mini-batch
  - Train logistic regression and update weights at every step

- ▷ **Improved our score but not significantly**

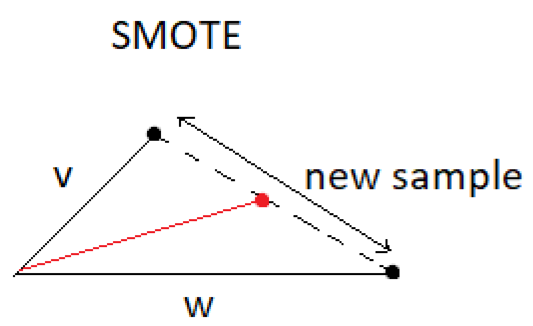
Oversampling

▷ General idea

- Train a classifier on a bigger sample of the data
- The sample is balanced adding more outliers
- Dupplicate outliers for example
- Problem: Overfitting

▷ How to sample new anomalies

- SMOTE (Synthetic Minority Over-Sampling Technique)



▷ Improved our score a lot

Our final submission

▷ Voting Classifier

- When you have some classifiers that work well
- A way to combine them to balance their strengths and weaknesses
- Black-Box…

▷ Classifiers that voted

- Adaboost
- Random Forest
- Gaussian Mixture
- Logistic Regression

▷ Up to 89% on the test set

3.

Unsupervised Learning

A few ideas

Isolation Forest

- General idea
  - Only for continuous features
  - Tries to isolate anomalies
  - Works well if anomalies are very different from those of normal instances
  - Builds an ensemble of trees and anomalies are instances which have short average path length
- We just have two continuous features
- Anomalies are not that different for these two features
- Maybe Veolia has more continuous features

4.

Metrics

ROC-AUC and PRECISION-RECALL-AUC

ROC-AUC weakness for imbalanced data

▷ Large number change in the False Positive Rate just leads to a small change in the ROC

▷ Overestimating the risk of failure is not penalized

▷ Example for 2 points in ROC/PR space:

- ❑ 10000 pipes and 100 will fail in 2014

1. Classifier 1 predicts 100 risky pipes with 90 True Positives
2. Classifier 2 predicts 500 risky pipes with 90 True Positives

- ❑ ROC:

1. Classifier 1: 0.9 True Positive Rate and 10/10000=0.001 False Positive Rate
2. Classifier 2: 0.9 True Positive Rate and 410/10000=0.041 Flase Positive Rate

- ❑ Difference of only 0.040 : too small !

Precision-Recall AUC

- Precision against Recall area under the curve
- Precision = TP / TP + FN
- Recall = True Positive Rate
- Number of True Negative has no impact

- Back to our example:
  - 10000 pipes and 100 will fail in 2014
  1. Classifier 1 predicts 100 risky pipes with 90 True Positives
  2. Classifier 2 predicts 500 risky pipes with 90 True Positives

  - Precision-Recall AUC:
  1. Classifier 1: 0.9 Recall and 90/100=0.9 Precision
  2. Classifier 2: 0.9 Recall and 90/500=0.18 Precision

  - Difference of 0.72 : very significant !

4.

Conclusion

Conclusion

- ▷ Hard problem
- ▷ Recent and ongoing research on the subject
- ▷ Could be a good idea to try more Unsupervised Learning
- ▷ Choice of the metric is very important and Veolia should try other metrics

Thanks!
Any questions?

adrien.savary@outlook.com