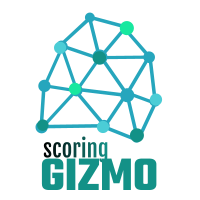
|  |
| --- |
| Name of the Score |
| Name of the Author, Department of the Author  **Description of the study** |
|  |
| |  |  | | --- | --- | | ***Entity name*** |  | | ***Study name*** |  | | ***Application*** |  | | ***Author*** |  | | ***Sponsor*** |  | | ***Recipients*** |  | | ***For information*** |  | | ***Operational Validation*** |  | | ***Domain / Field*** |  | | ***Train session*** | TRAIN\_bg\_stage2\_2024-07-30 10:38:21.777707\_no\_tag | |

20/11/2020



This score was built by SSC Analytics using Scoring Gizmo package:



|  |  |
| --- | --- |
| ***Period of study*** | 2017-01 – 2020-03 |
| ***Period of train*** | 2017-01 – 2019-12 (70%) |
| ***Period of test*** | 2017-03 – 2019-11 (30%) |
| ***Period of temporal validation*** | 2020-01, 2020-02, 2020-03 |
| ***Excluded periods*** | [] |
| ***Data origin*** |  |
| ***Statistical tools*** | Python, Sklearn (Random Forest, Decision tree), Xgboost, Logistic regression |
| ***Structural validation method*** | Random split of building sample (70/30 ratio) |
| ***Criterion*** |  |
| ***Performances*** |  |
| ***Roles*** |  |

Content

[1. Presentation of the situation 3](#__RefHeading___Toc2203_424922862)

[1.1 The Overview 3](#__RefHeading___Toc2205_424922862)

[2. Presentation and analysis of project 3](#__RefHeading___Toc2207_424922862)

[2.1 The perimeter of the study 3](#__RefHeading___Toc2209_424922862)

[2.1.1 Criterion study 3](#__RefHeading___Toc2211_424922862)

[2.2 The wherewithal 4](#__RefHeading___Toc2213_424922862)

[2.3 Target variable 4](#__RefHeading___Toc2215_424922862)

[3. Modelling 4](#__RefHeading___Toc2217_424922862)

[3.1 Statistical method 4](#__RefHeading___Toc2219_424922862)

[3.2 Candidate variables 5](#__RefHeading___Toc2221_424922862)

[3.3 Model selection 5](#__RefHeading___Toc2223_424922862)

[3.4 Variables selection 5](#__RefHeading___Toc2225_424922862)

[4. Final model 7](#__RefHeading___Toc2227_424922862)

[4.1 Segments stability 8](#__RefHeading___Toc2229_424922862)

[4.2 Variables stability 10](#__RefHeading___Toc2231_424922862)

[4.3 Score grid 14](#__RefHeading___Toc2233_424922862)

[4.4 Distribution of the deciles on valadation population 15](#__RefHeading___Toc2235_424922862)

[1.1 Accumulation Points by probability 15](#__RefHeading___Toc2237_424922862)

[1.1 Uplift curve and scoring bands 16](#__RefHeading___Toc2239_424922862)

[1.2 Efficiency metrics 16](#__RefHeading___Toc2241_424922862)

[2. Implementation of the score 17](#__RefHeading___Toc2243_424922862)

[2.1 Principle 17](#__RefHeading___Toc2245_424922862)

[2.2 Binning of variables 17](#__RefHeading___Toc2247_424922862)

[2.3 Description of the implementation strategy 18](#__RefHeading___Toc2249_424922862)

[2.4 Follow up in production & real efficiency measures 18](#__RefHeading___Toc2251_424922862)

[3. Annex 18](#__RefHeading___Toc2253_424922862)

[3.1 Description of the data 18](#__RefHeading___Toc2255_424922862)

# Presentation of the situation

## The Overview

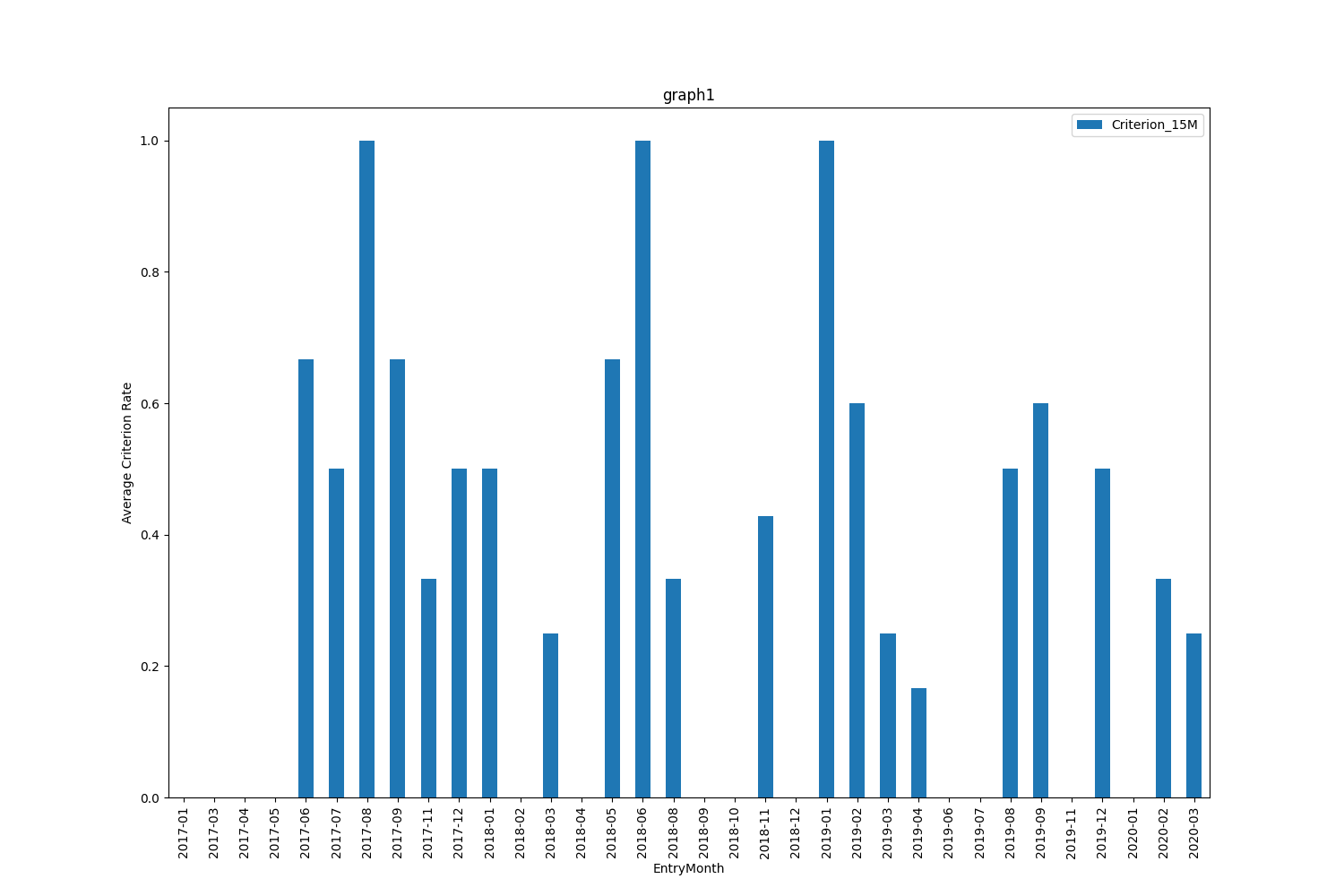
*Fill in this part during the scoping definition of the project.*

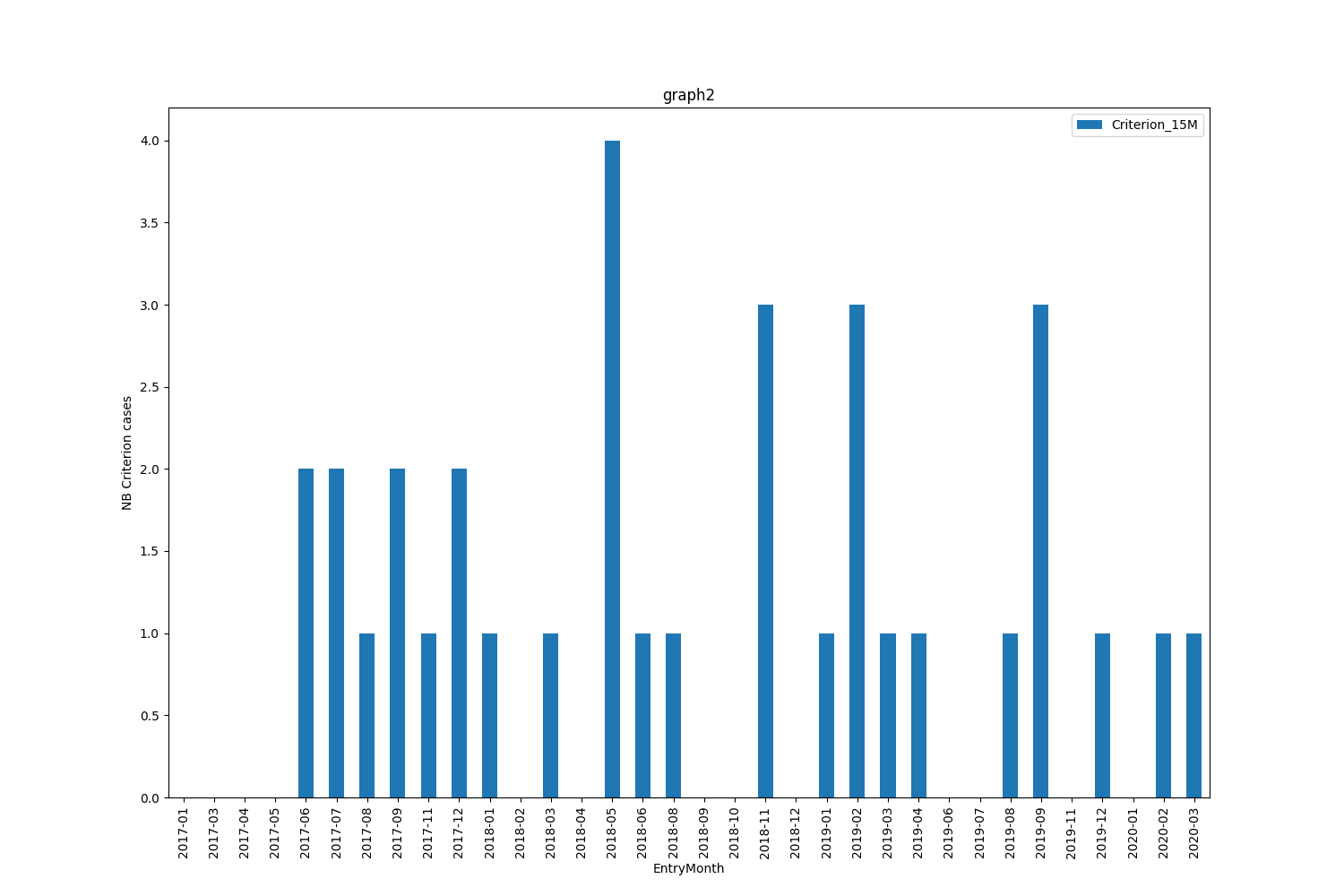
# Presentation and analysis of project

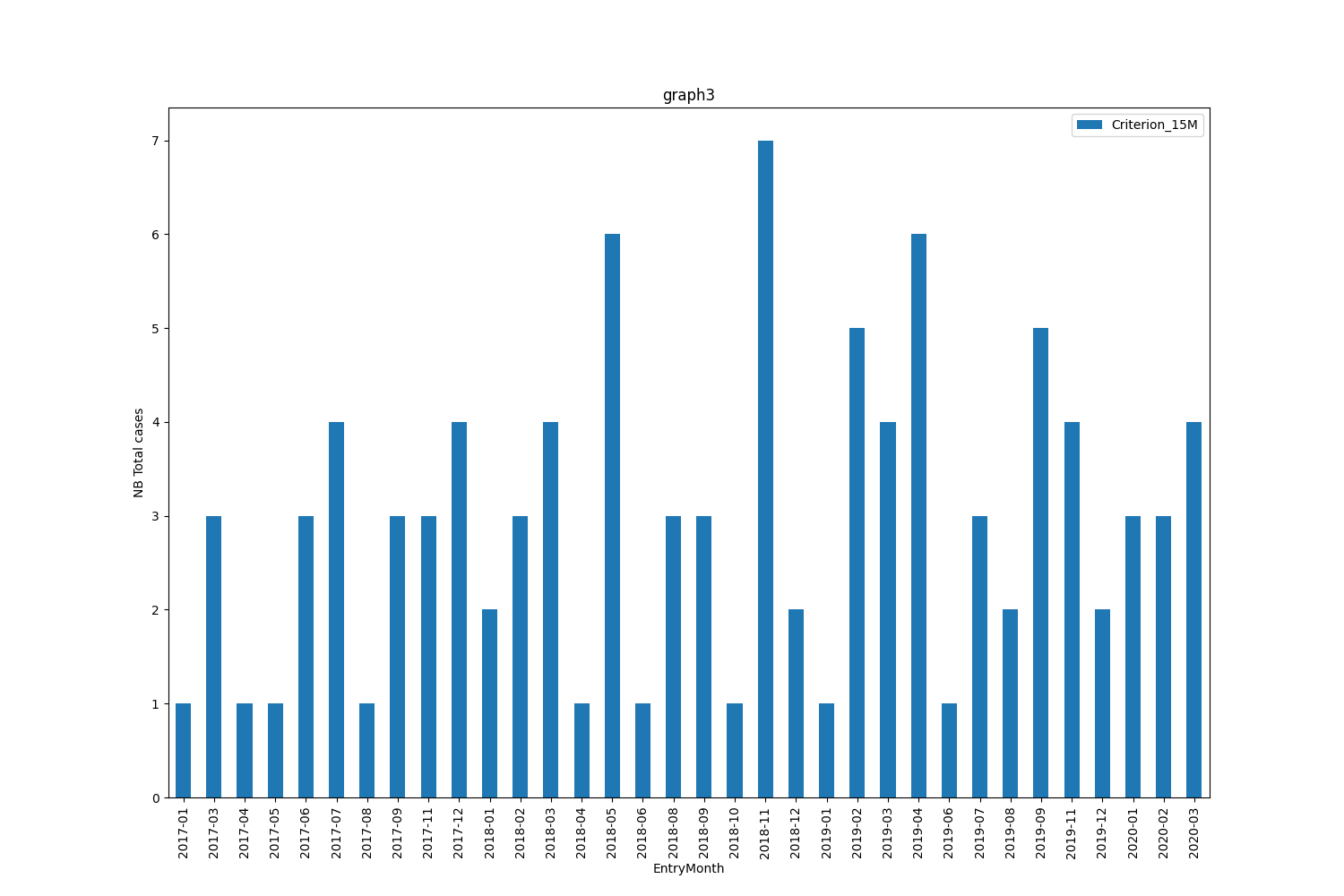
## The perimeter of the study

*Fill in this part during the scoping definition of the project.*

### Criterion study – the whole data set before removing periods

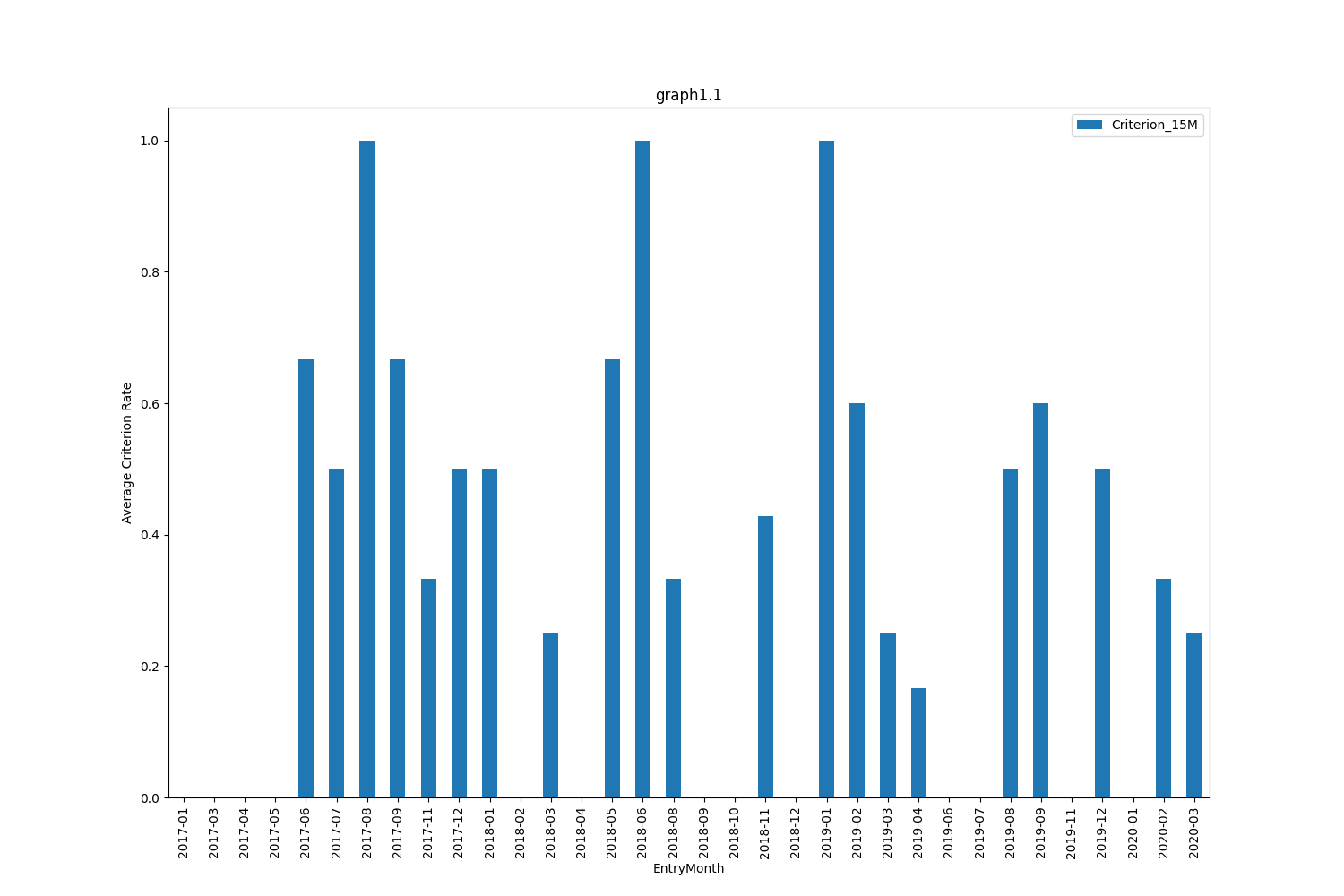


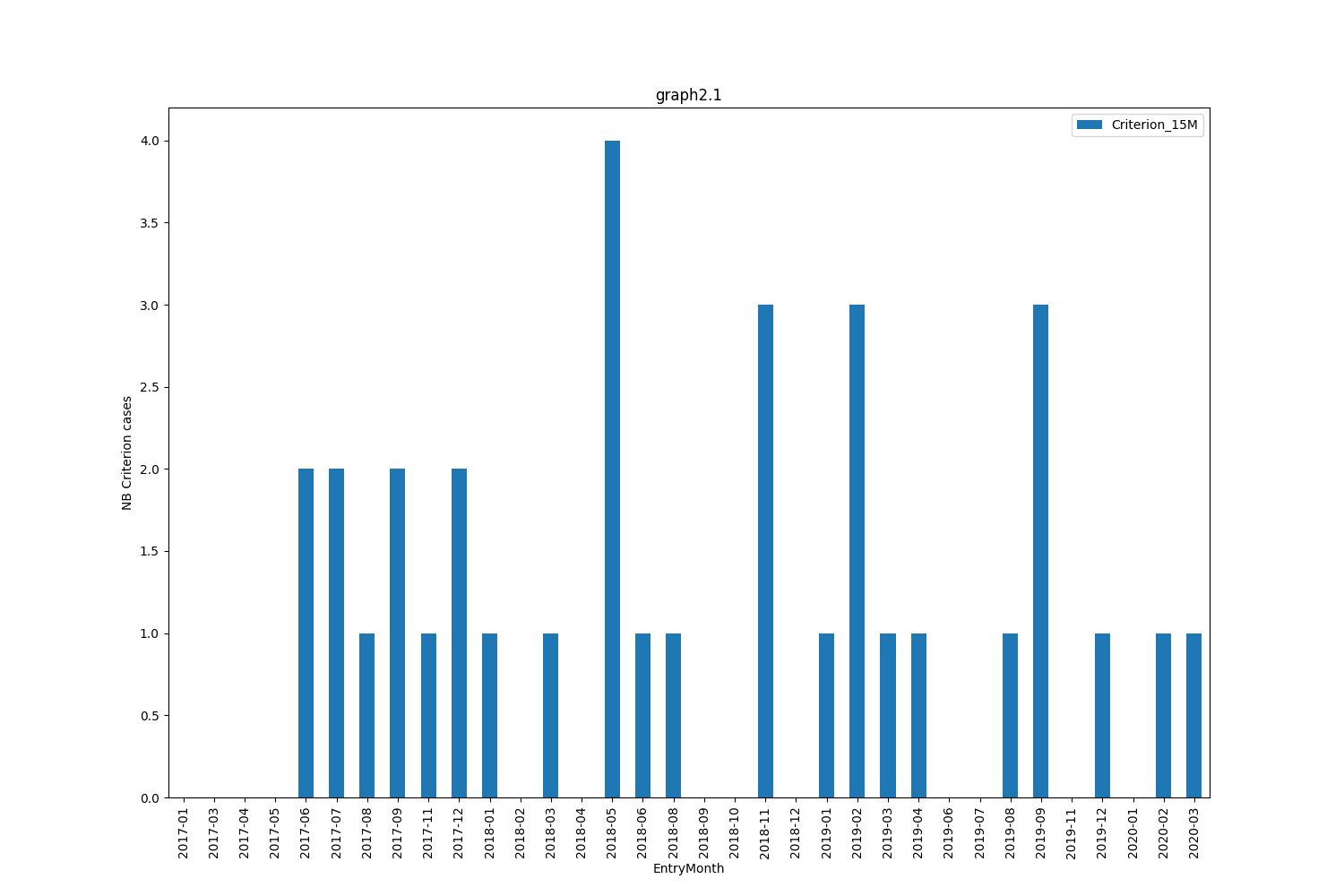


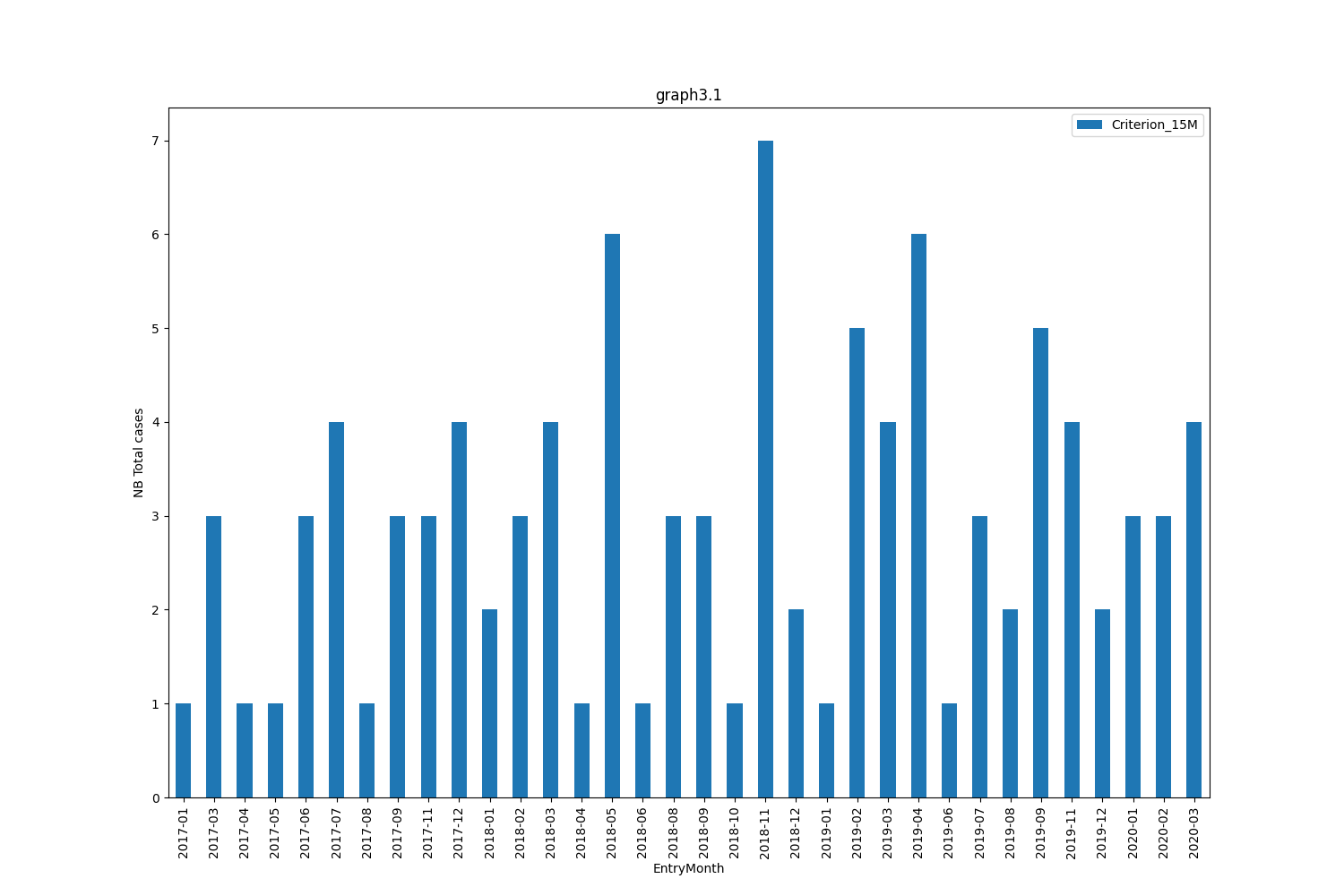


*Insert your comments here*

### Criterion study – the after removing periods







*Insert your comments here*

## The wherewithal

Dataset was built in SQL. The study done in Python.

## Target variable

*Fill in during the scoping phase.*

# Modelling

## Statistical method

For the modelling part the dataset was split on 2 parts- Train (70%) and Test (30%) + 3 temporal validation samples.

The volumes are below:

* Train: 60 nb cases, with 0.367 criterion rate
* Test: 30 nb cases, with 0.333 criterion rate
* t1: 3 nb cases, with 0.0 criterion rate
* t2: 3 nb cases, with 0.333 criterion rate
* t3: 4 nb cases, with 0.25 criterion rate

## Candidate variables

*Describe the source of the data.*

**The description of the dataset is in appendix**

## Model selection

Couple of different statistical models were used in competition to be able to pick the best option from (1) efficiency point of view, but also considering the technical and business implementations.

Models used in competition: XGBoost, Random forest, Logistic regression, Decision tree

## Variables selection

All variables were with treated missing files and after a feature engineering process was performed in order to enrich the predictors in the data set.

At the end, all features with correlation greater than 5% were used for modelling.

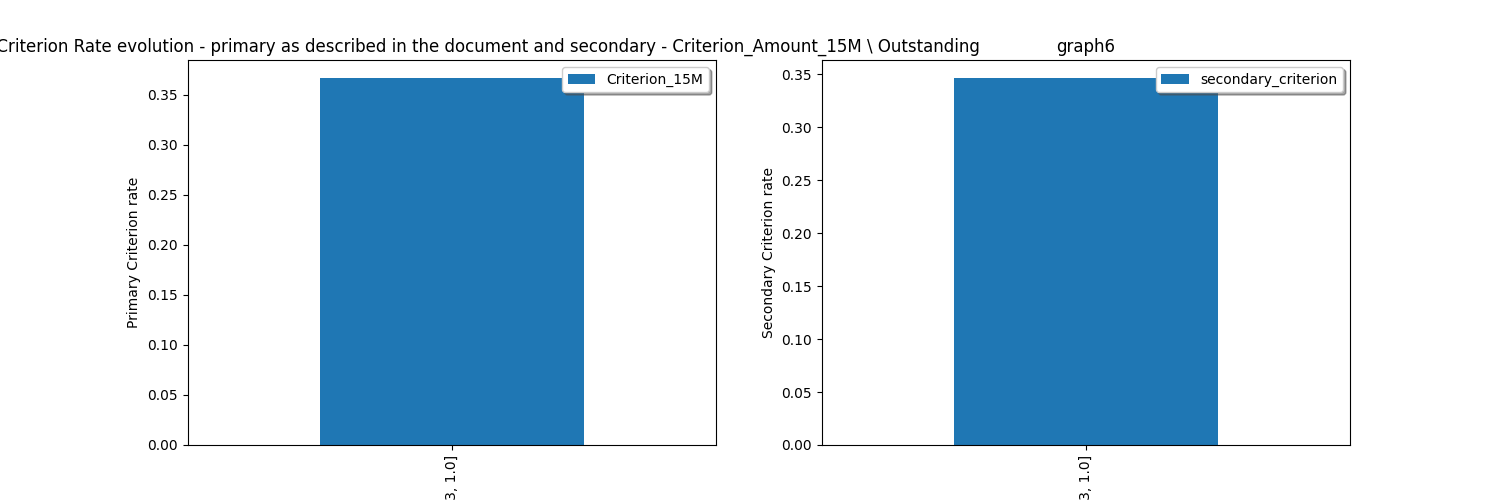
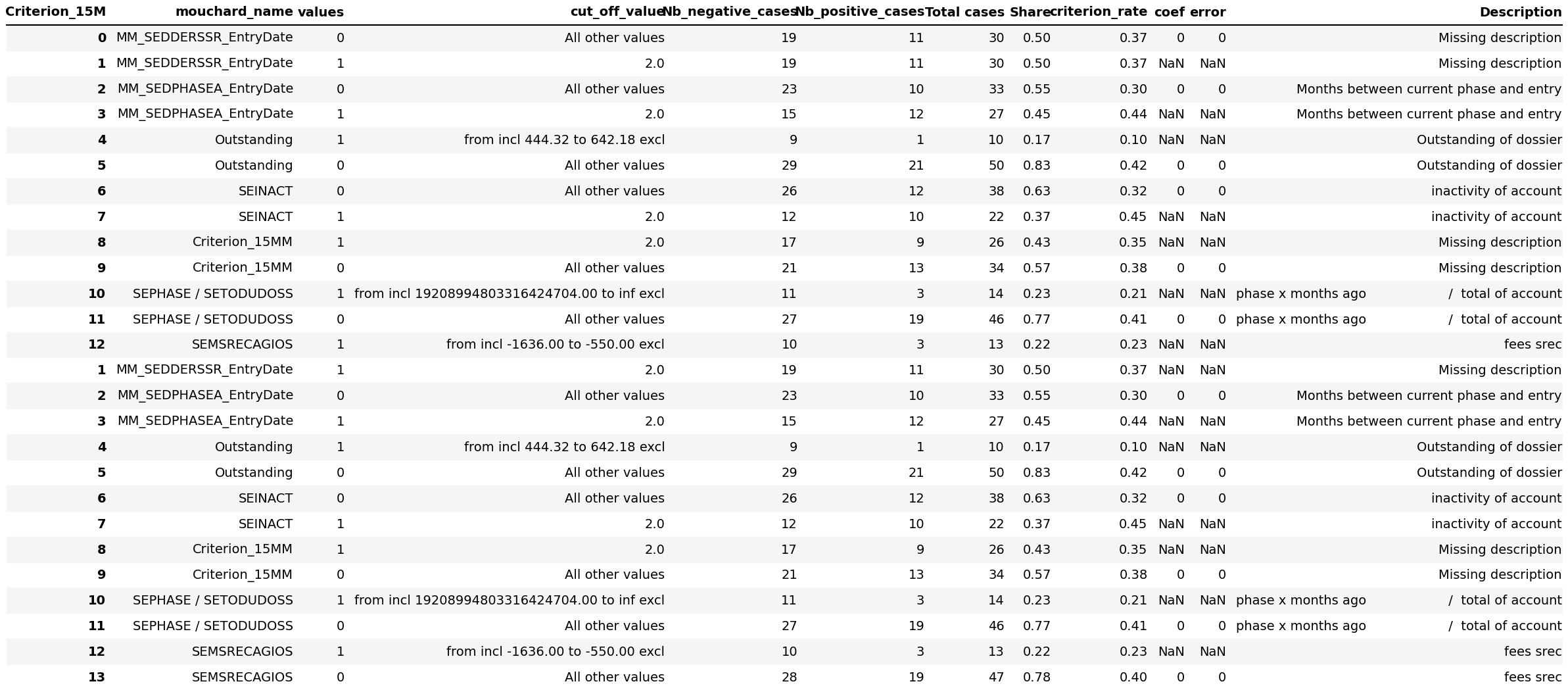
Share of missing values by variable is in the appendix.

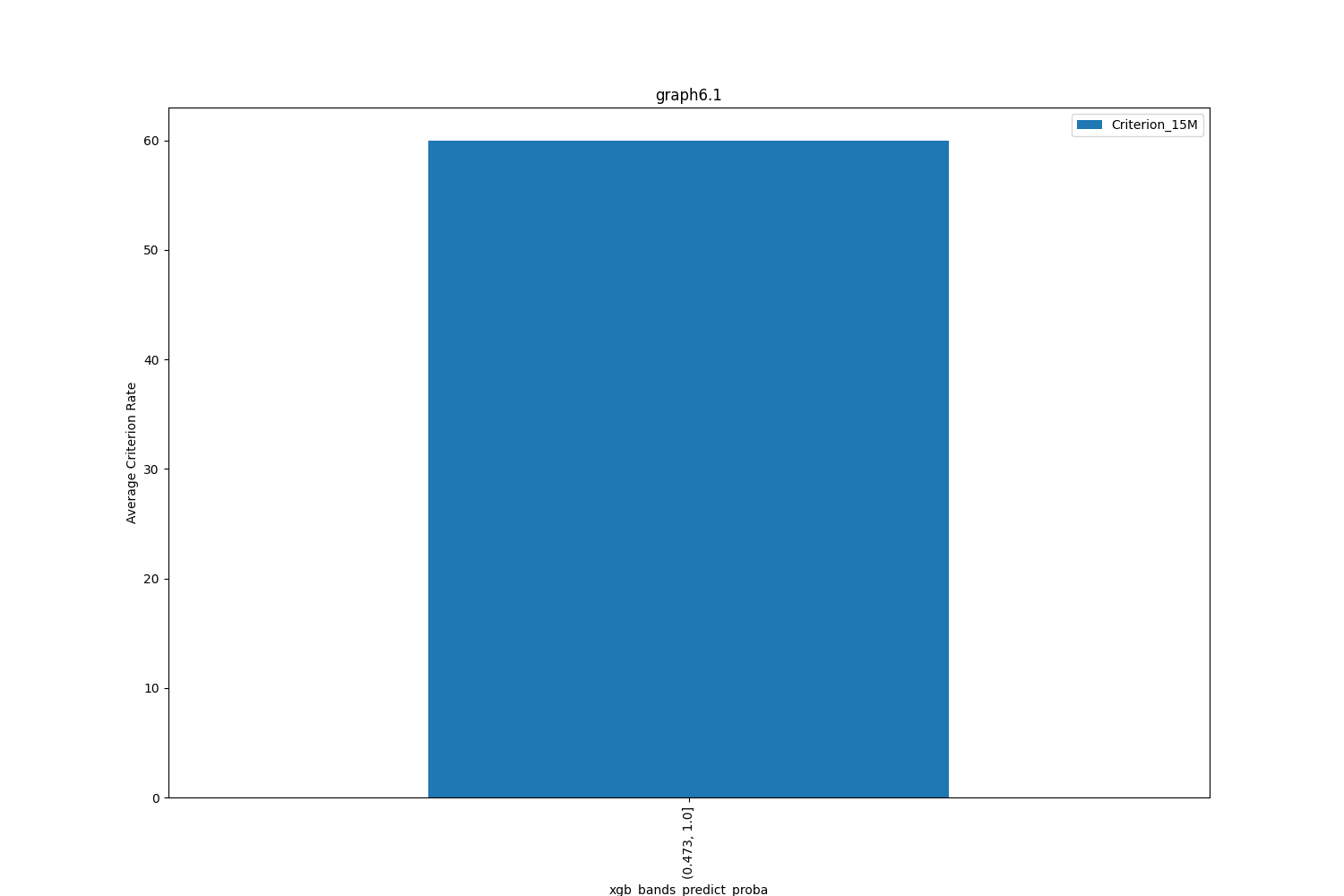
Missing values were treated: 1

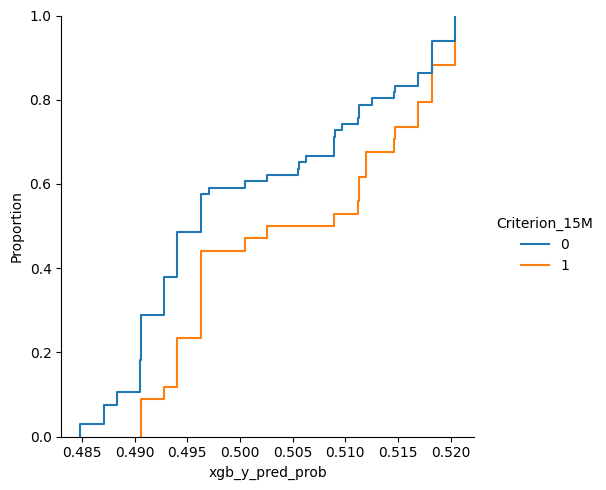
# Final model

The chosen model is: XGBoost Classifier

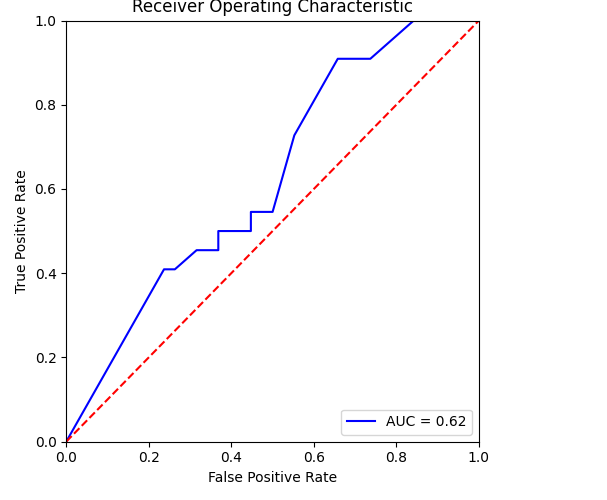
Based on the nb of probabilities 21 the following 2 bands (cut-offs) are possible:



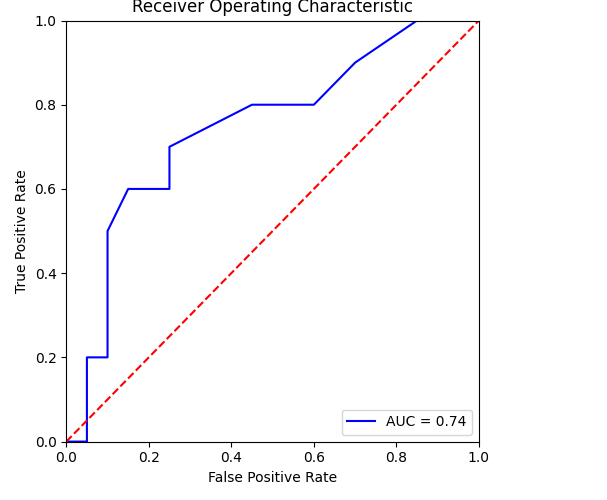




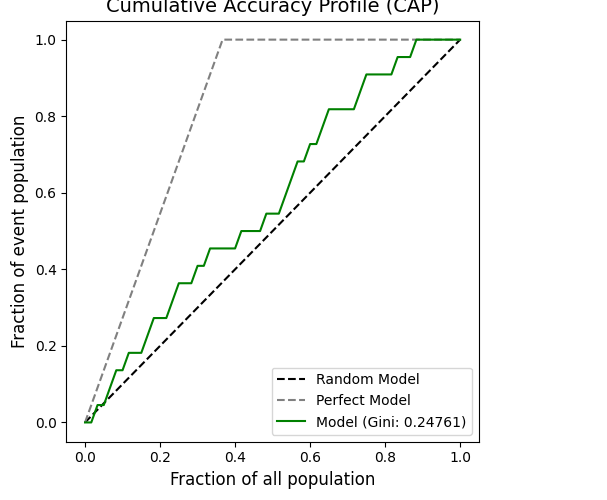
**ROC Curve Train**



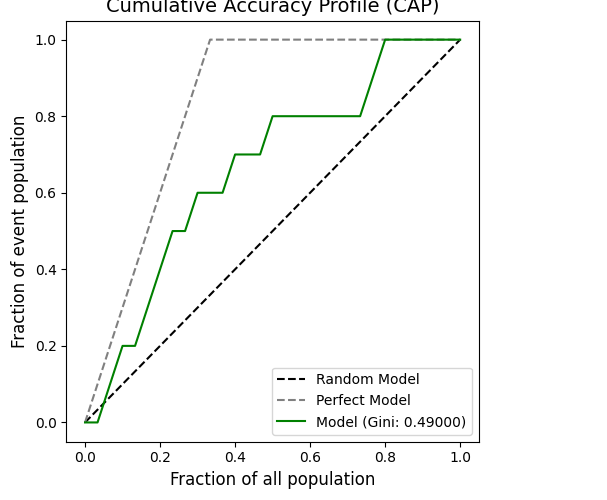
**ROC Curve Test**



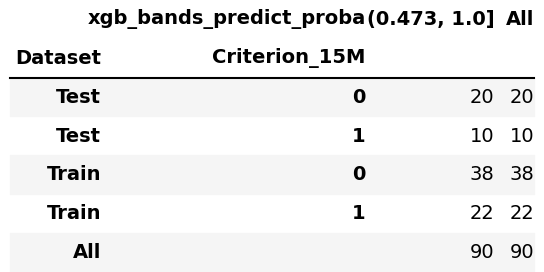
**CAP** curve for Train sample:



**CAP** curve for Test sample:

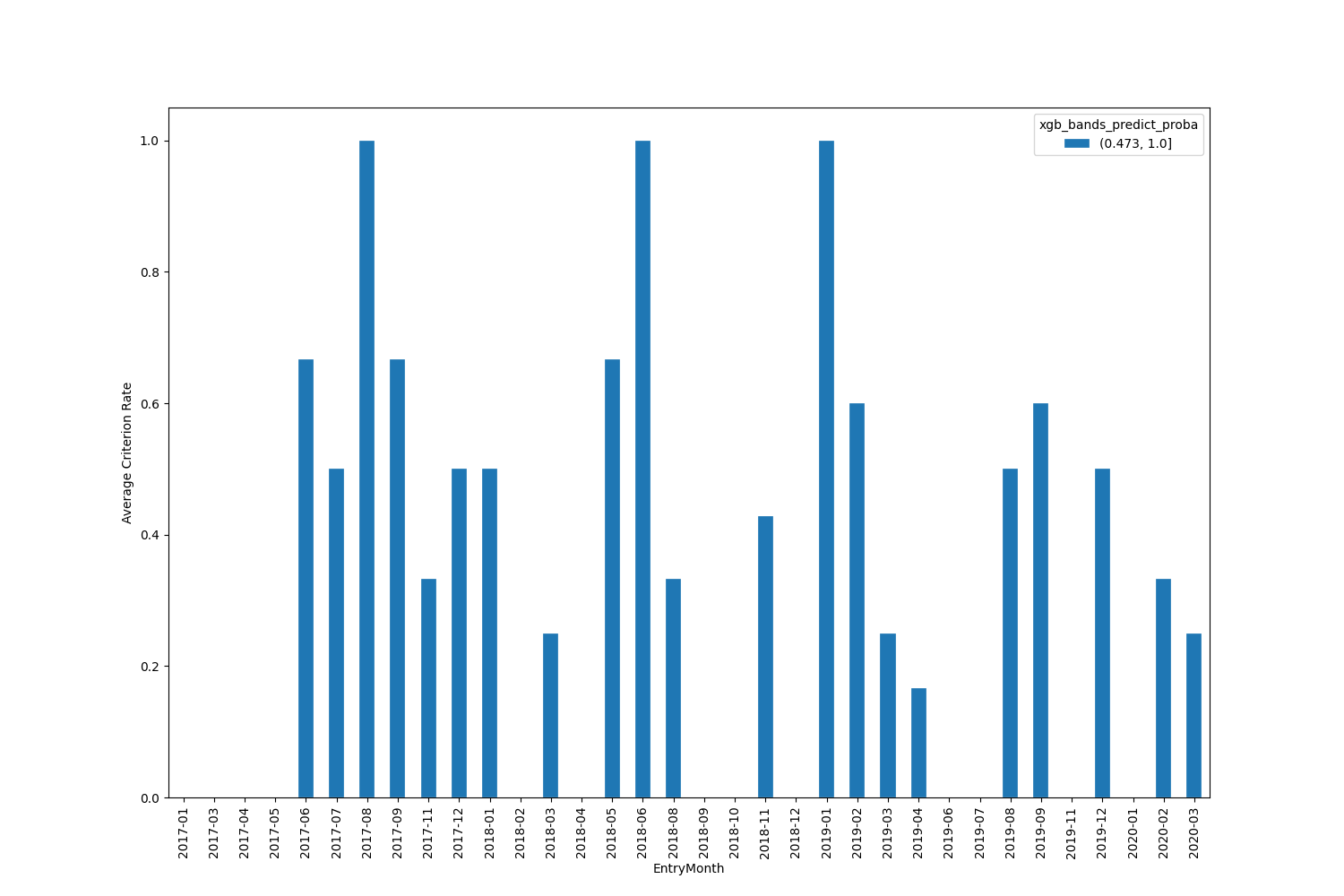


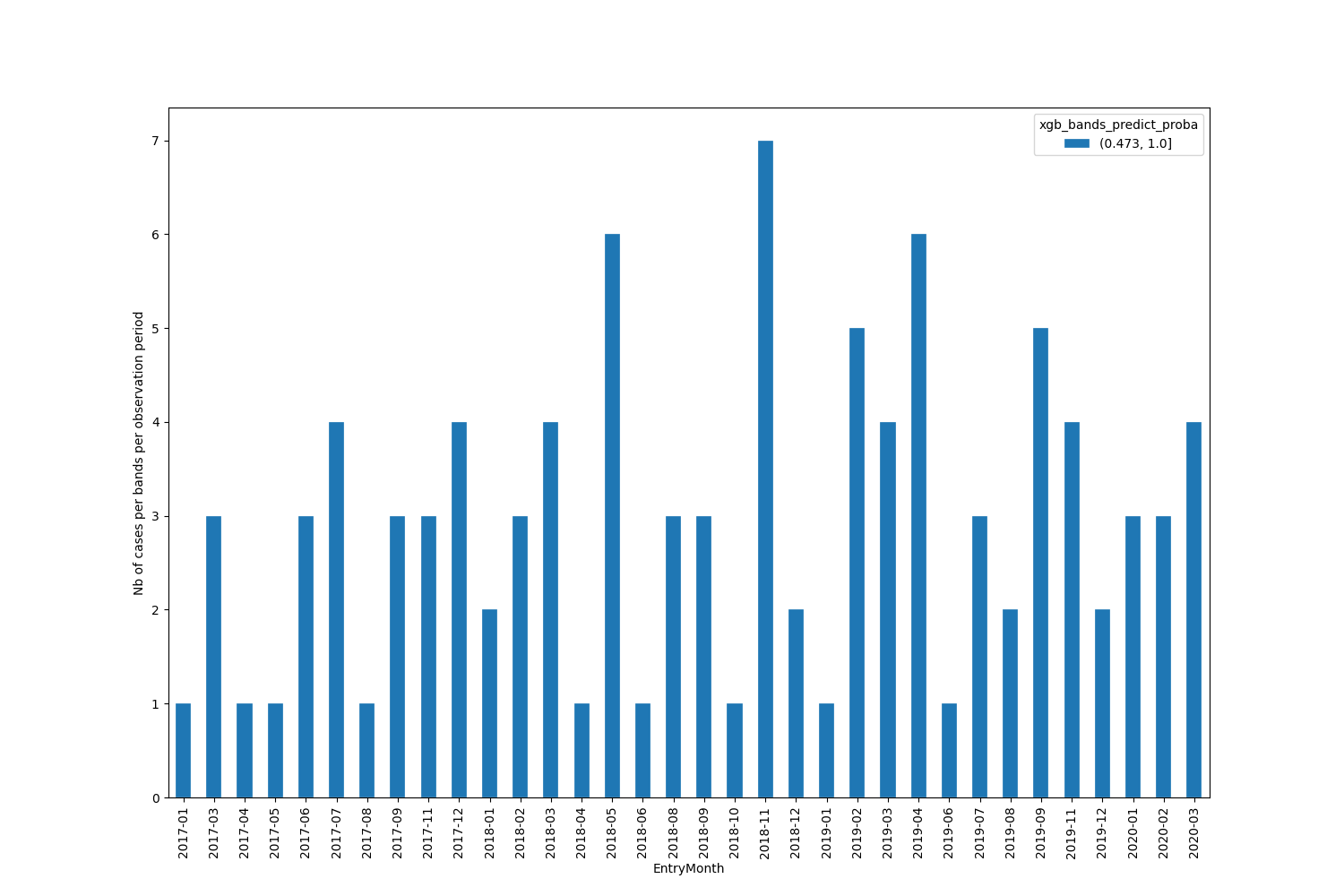
The split between bands for Train and Test with positive and negative criterion:

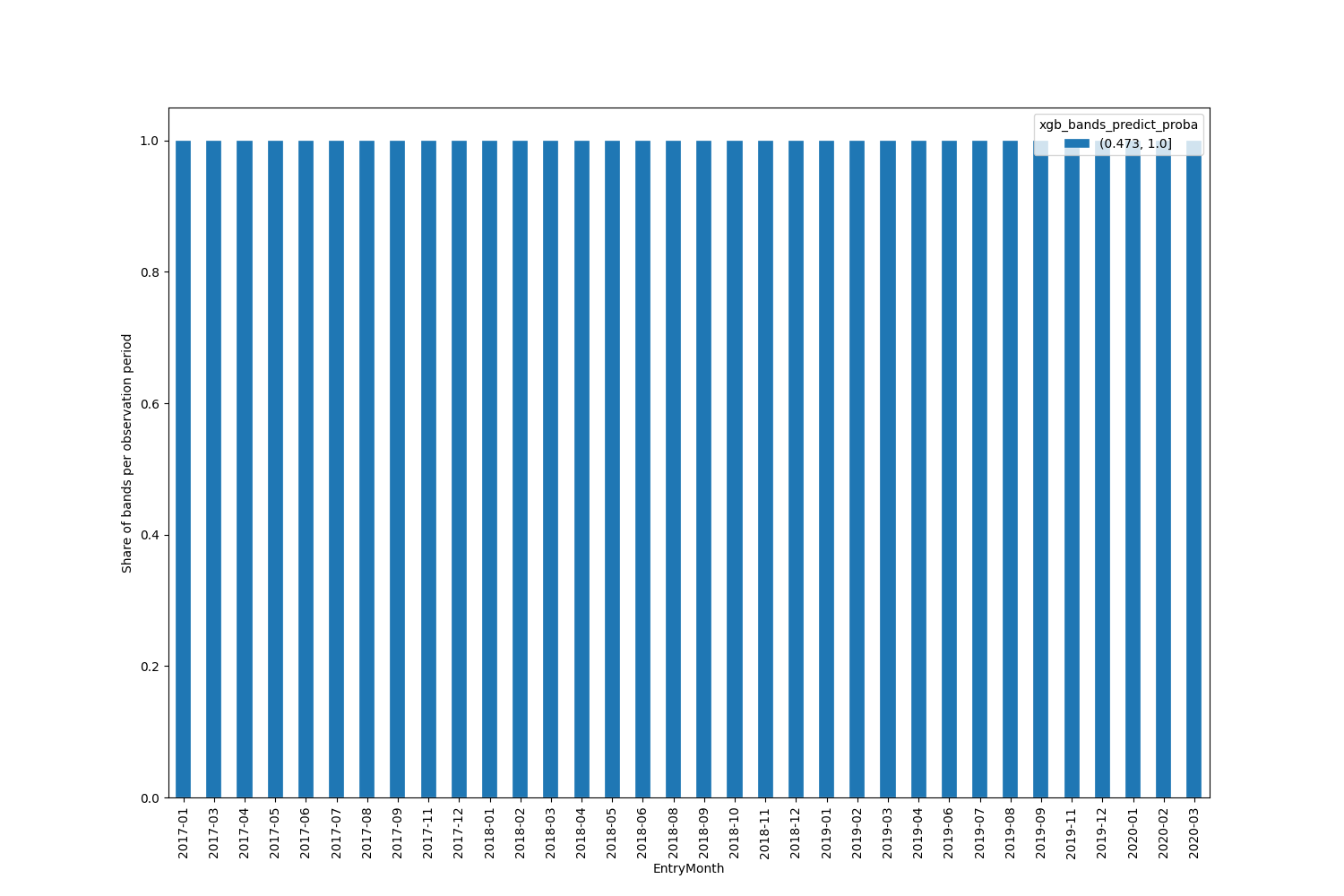


## Segments stability

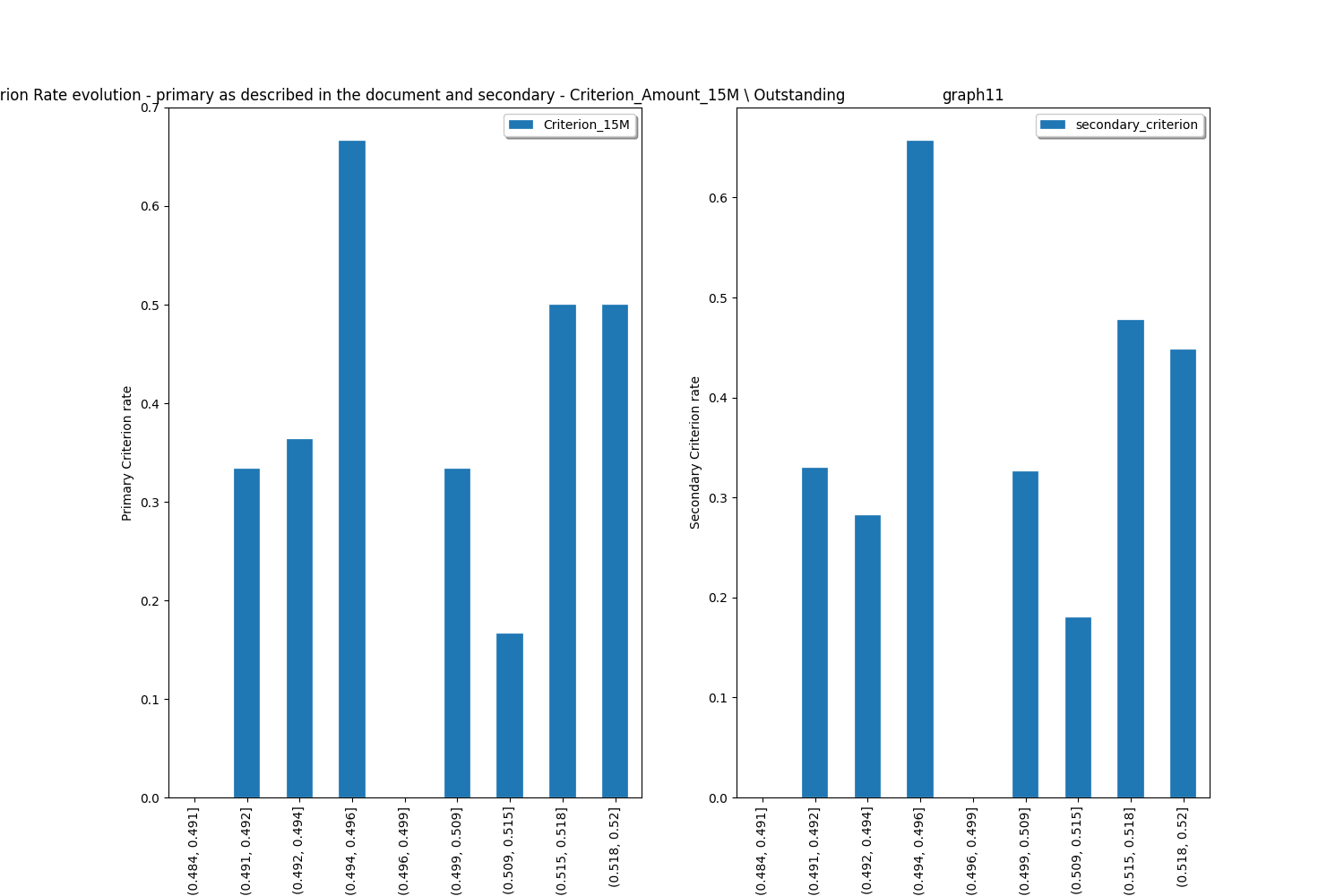
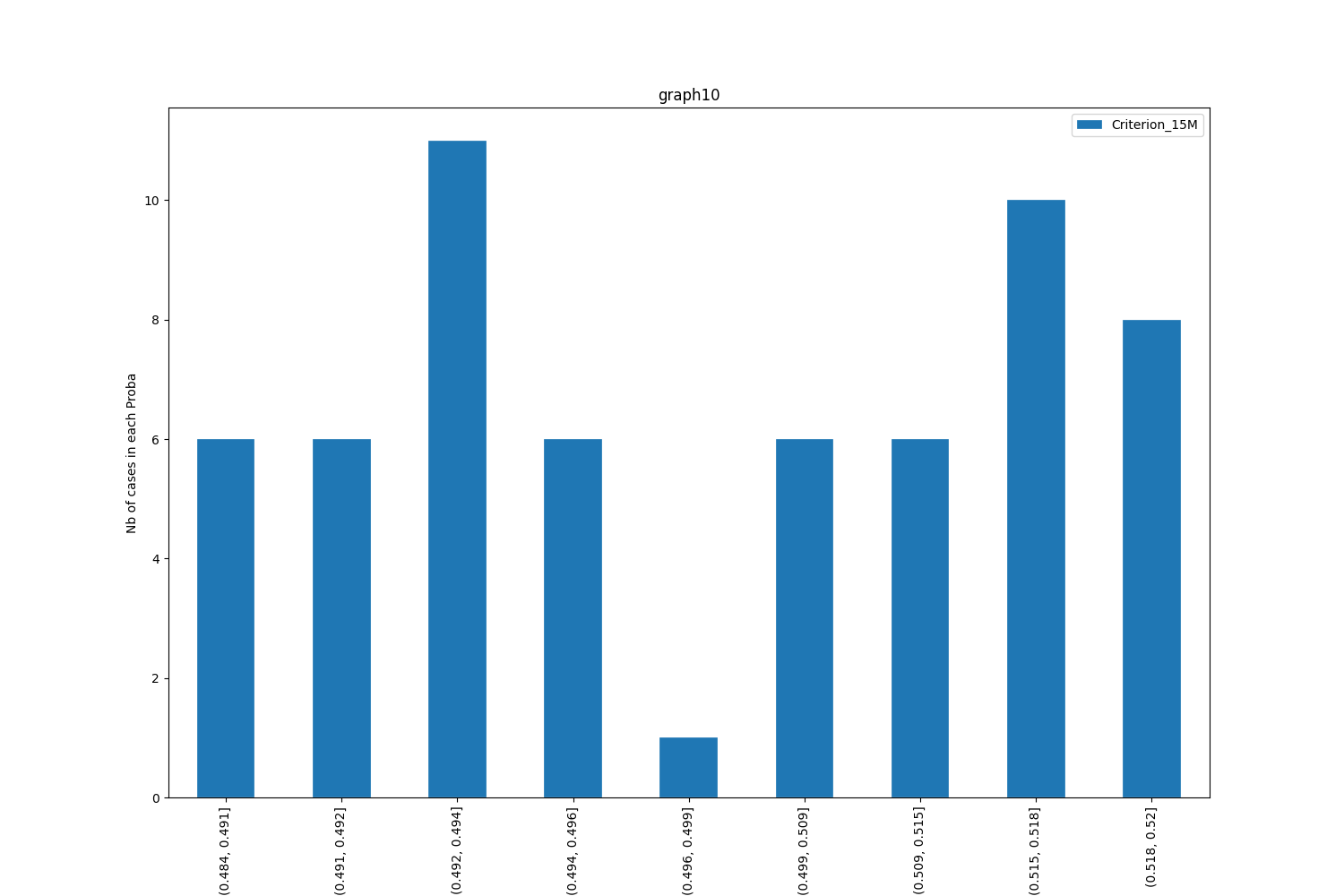
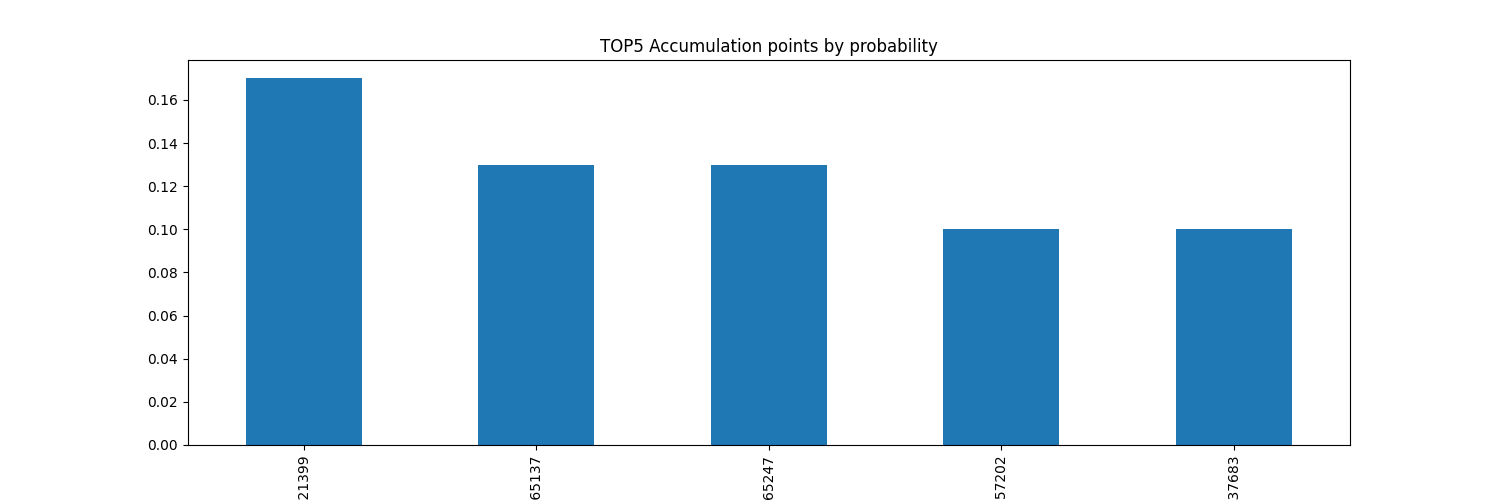
*Evolution of Positive and Negative Criterion by bands and observation points for all periods in Train, Test, T1, T2 and T3:*



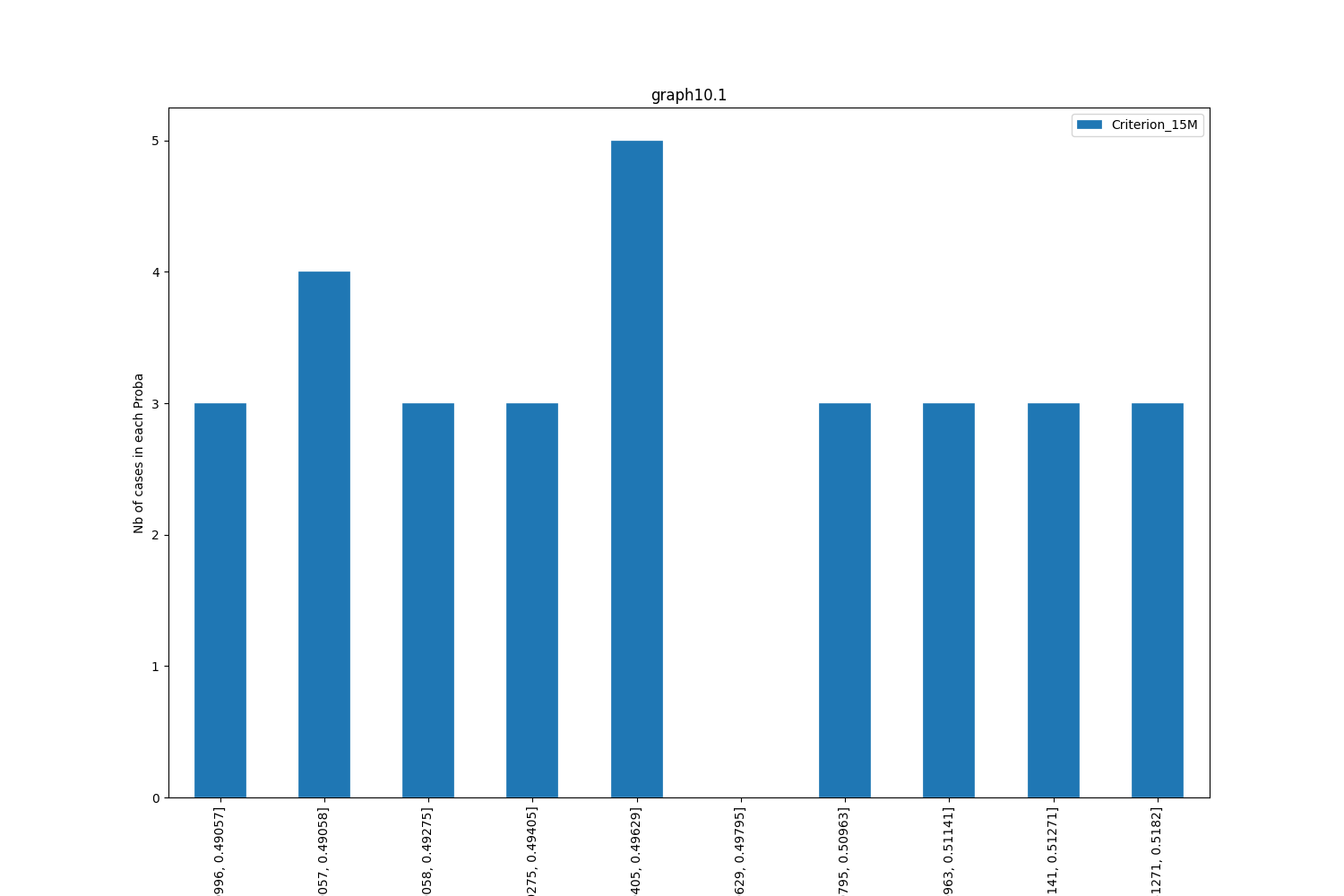
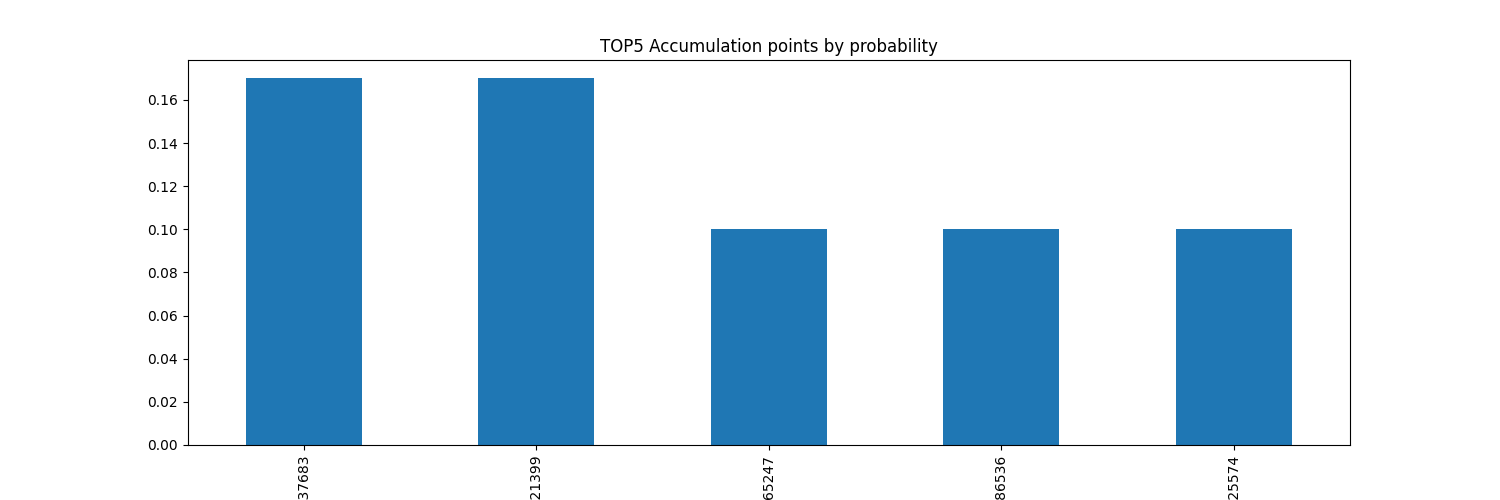


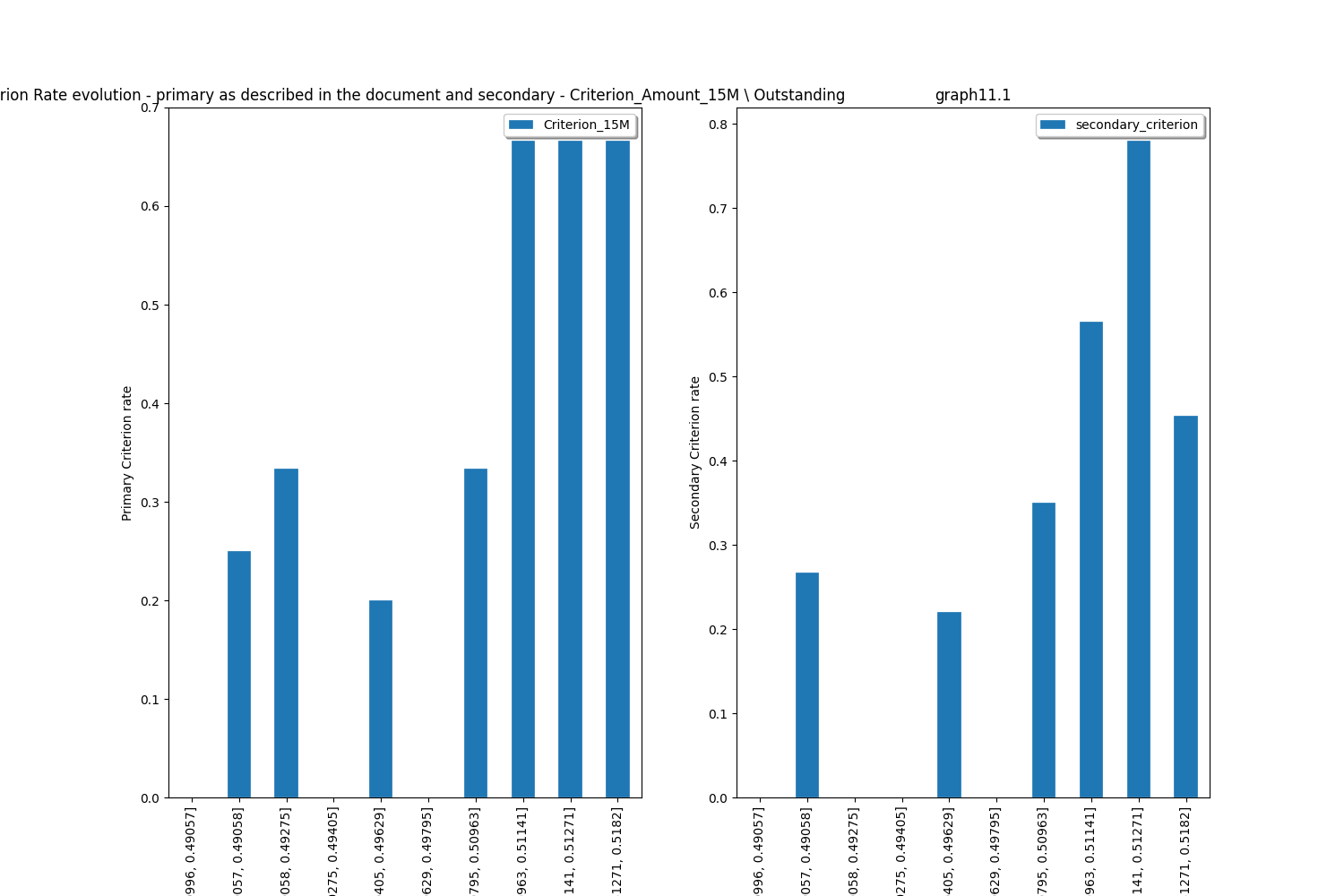


*For* ***Train*** *set in the graphs below are shown volumes the deciles, split by Proba, and the evolution of the criterion rate by decile:*



*For* ***Test*** *set in the graphs below are shown volumes the deciles, split by Proba, and the evolution of the criterion rate by decile:*

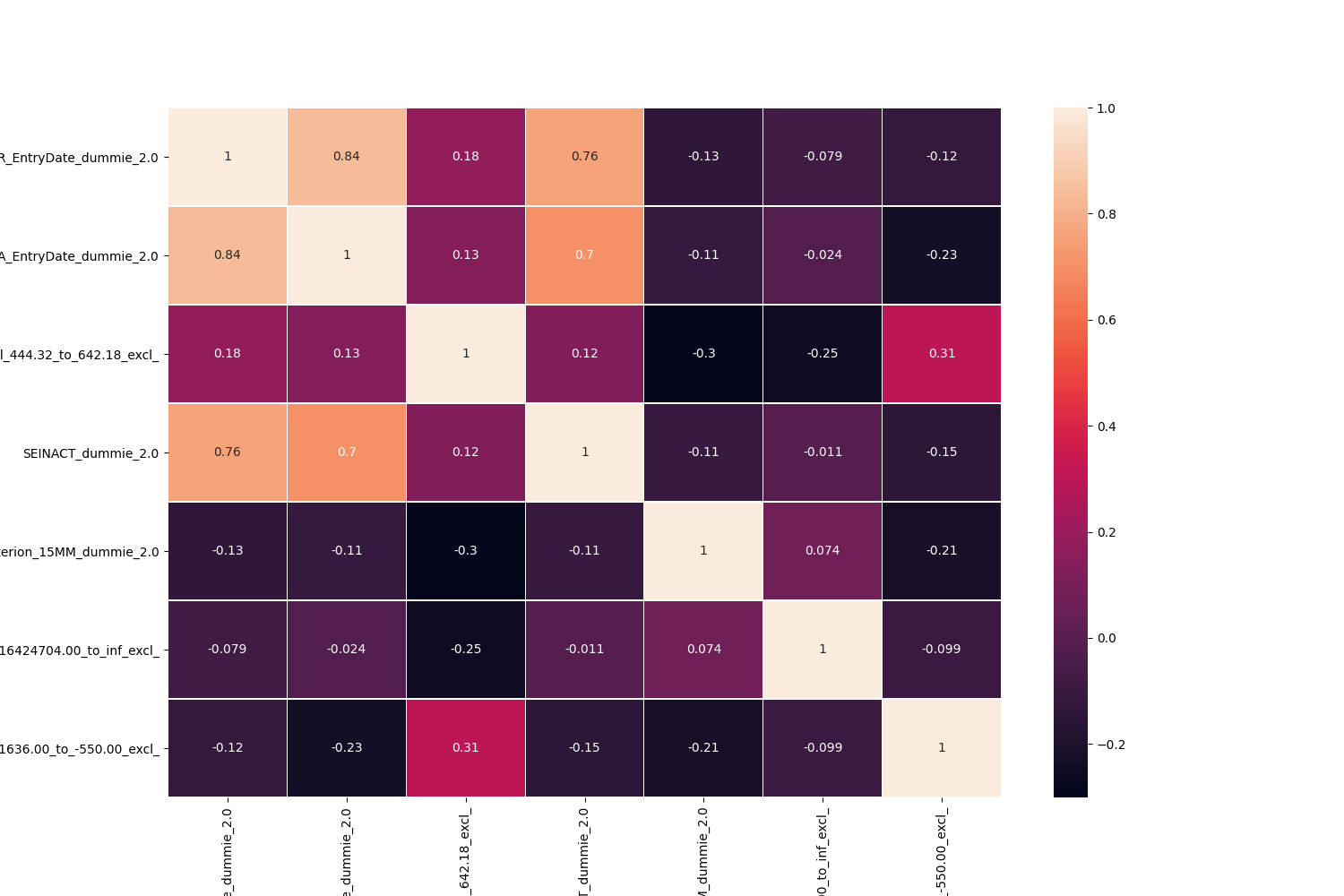


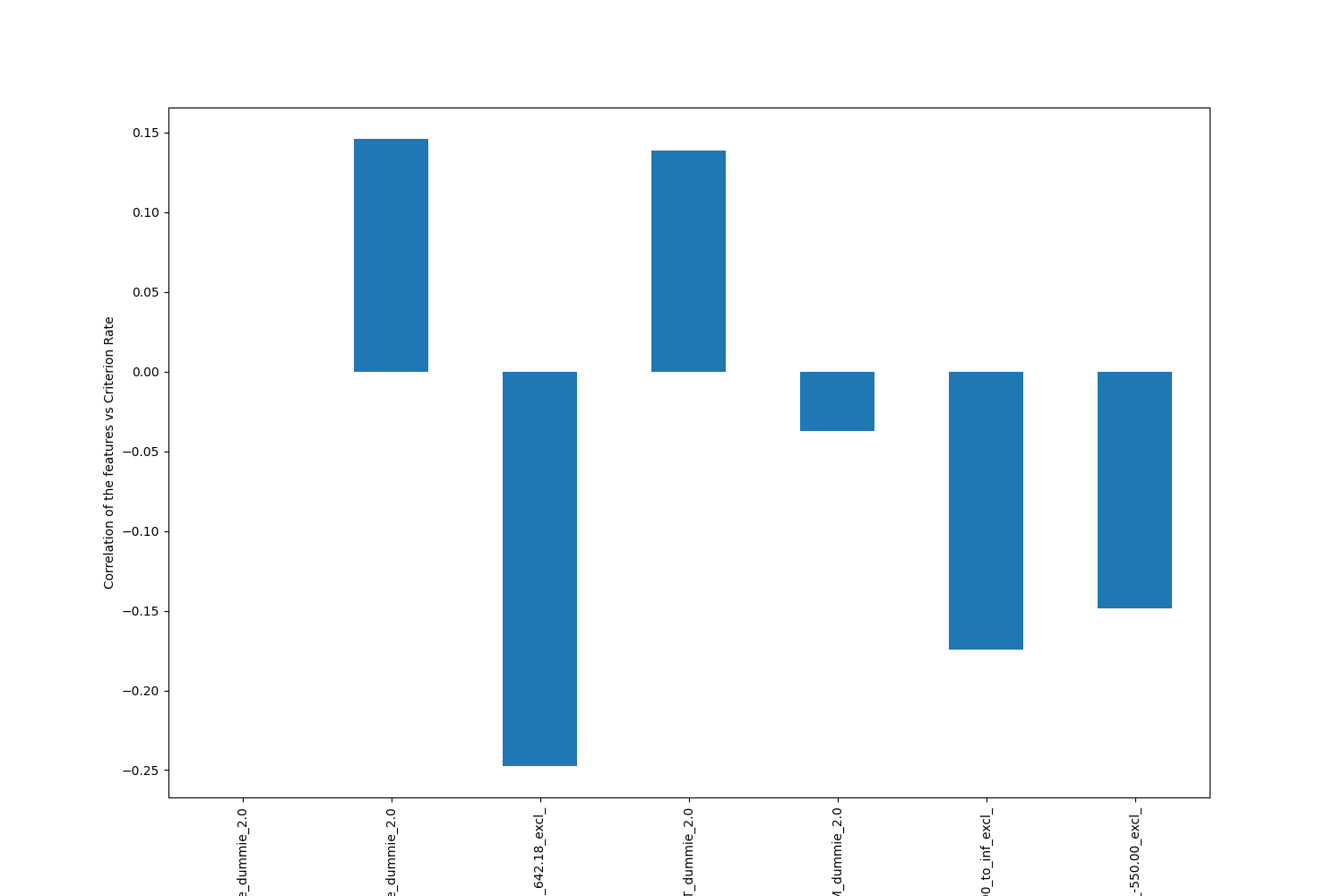


## Variables stability

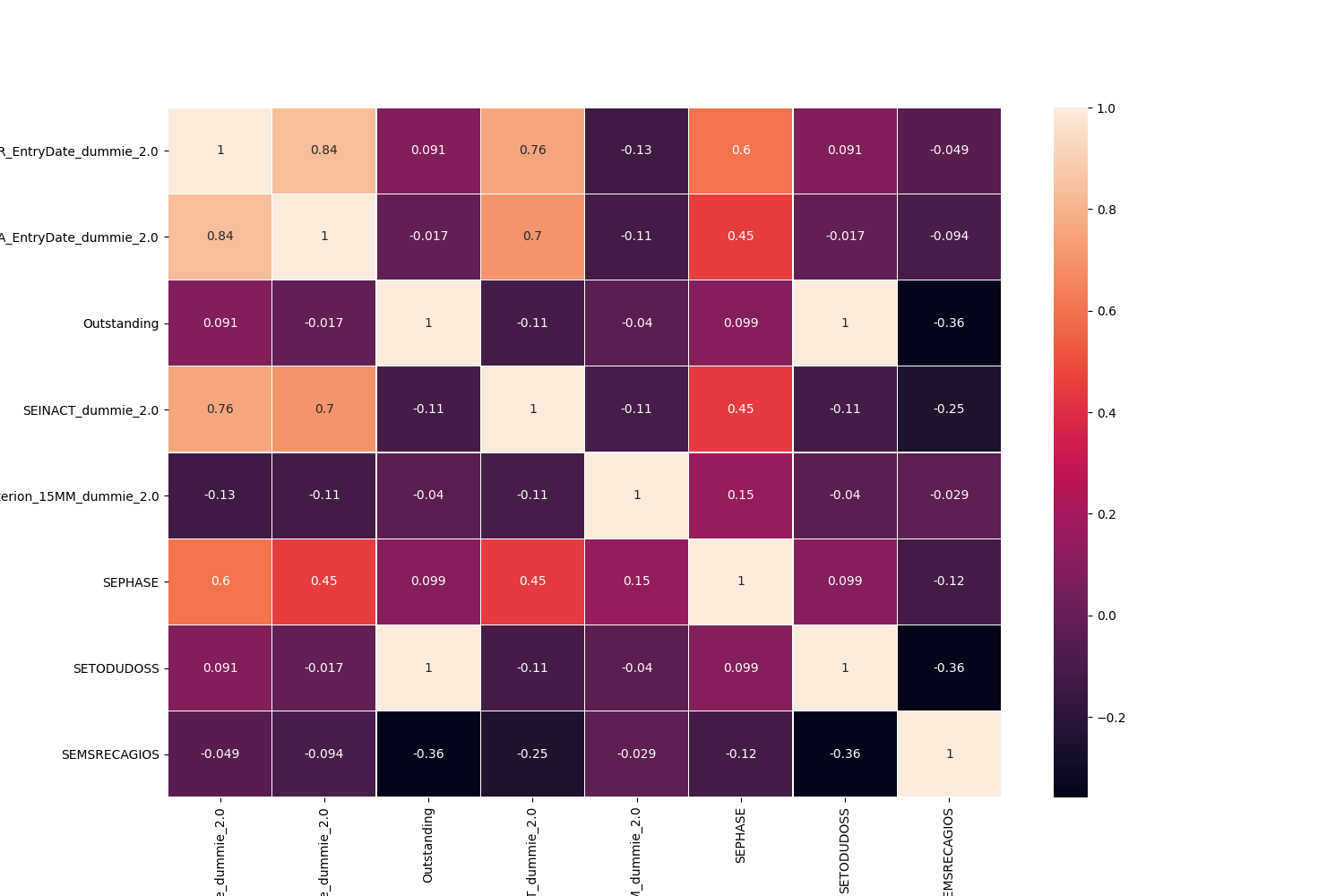
Pierson`s correlation between final and raw features used for the final model can be observed on the graph below.

List of final features: ['Criterion\_15MM\_dummie\_2.0' 'MM\_SEDDERSSR\_EntryDate\_dummie\_2.0'  
 'MM\_SEDPHASEA\_EntryDate\_dummie\_2.0'  
 'Outstanding\_binned\_dummie\_from\_incl\_444.32\_to\_642.18\_excl\_'  
 'SEINACT\_dummie\_2.0'  
 'SEMSRECAGIOS\_binned\_dummie\_from\_incl\_-1636.00\_to\_-550.00\_excl\_'  
 'SEPHASE\_div\_ratio\_SETODUDOSS\_binned\_dummie\_from\_incl\_19208994803316424704.00\_to\_inf\_excl\_']





List of raw features: ['Criterion\_15MM\_dummie\_2.0' 'MM\_SEDDERSSR\_EntryDate\_dummie\_2.0'  
 'MM\_SEDPHASEA\_EntryDate\_dummie\_2.0' 'Outstanding' 'SEINACT\_dummie\_2.0'  
 'SEMSRECAGIOS' 'SEPHASE' 'SETODUDOSS']



# Implementation of the score

## Principle

To be decided

## Description of the implementation strategy

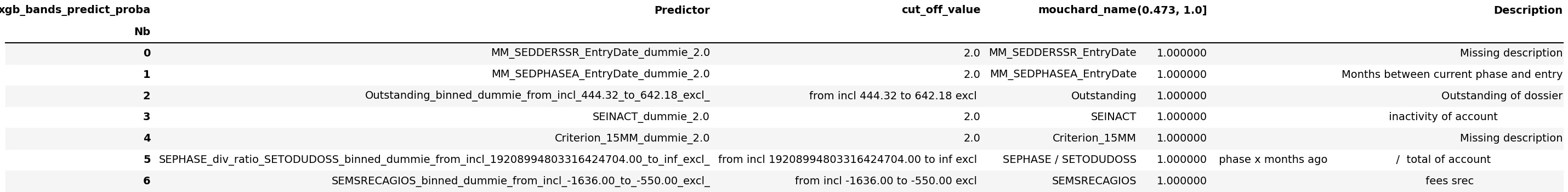
To be done

## Follow up in production & real efficiency measures

MUSE to be implemented

# Annex

# Description of the data



# Missing values



# Feature stability

