Polish Companies Bankruptcy Data

ALEXANDRE FORESTIER, JULIE PICOT

ESILV A4 – DIA 1

DATA SET INFORMATION

- Bankruptcy prediction of polish companies
  - Analyse in the period of 2000 – 2012 for the bankrupt companies
  - Analyse in the period of 2007 – 2013 for the still operating companies

- 5 classification cases about financial rates of the forecasting period and the corresponding class label that indicates bankruptpcy status

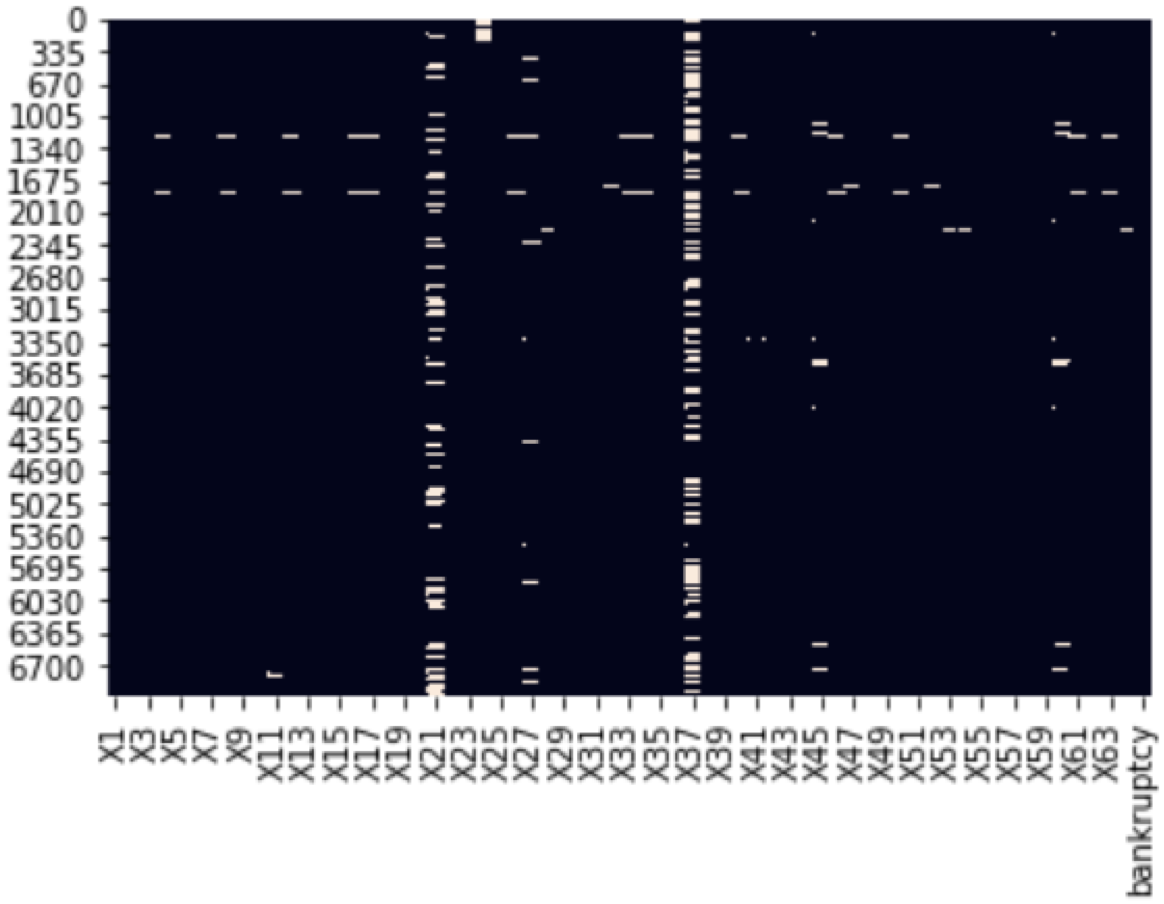
- This analysis allows companies to evaluate their situation. It is also useful for banks to decide whether or not to grant loans to them.

- The objective of our analysis is to find the best bankruptcy forecasting model and the potential anomalies that would affect the company's situation.

DATA EXPLORATION

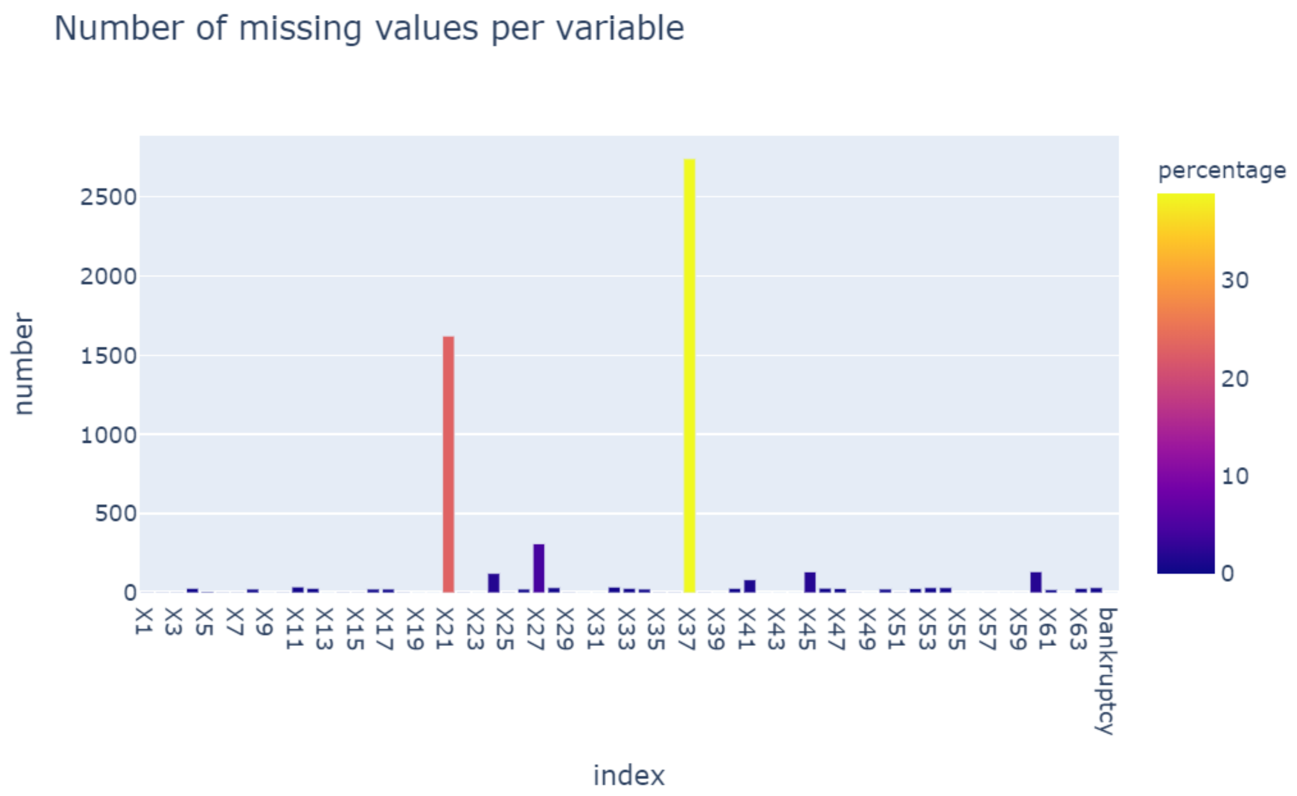
- After a first exploration of the data, we decided to load a shorter naming system for variables : X1, X2, X3…

- Then we plot an heatmap to identify if we can see missing data and yes, we have a lot missing data.



DATA PREPARATION : MISSING DATA

- We have a lot of missing data for some variables and we won't use a median or an average filing for missing values. There is an average of 1.5 missing values per element.



- To fill the missing values, we use a Linear Regression Model to predict them.

MODELING

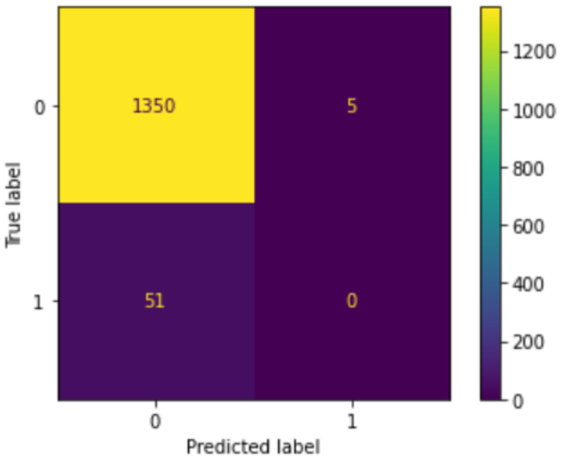
- We test 7 models and check for each the accuracy and confusion matrix :
  - Logistic regression
  - Decision Tree
  - Gaussian Naive Bayes
  - Random Forest Classifier
  - Boosting Classifier
  - KNN Classifier
  - Bagging Classifier

ACCURACY

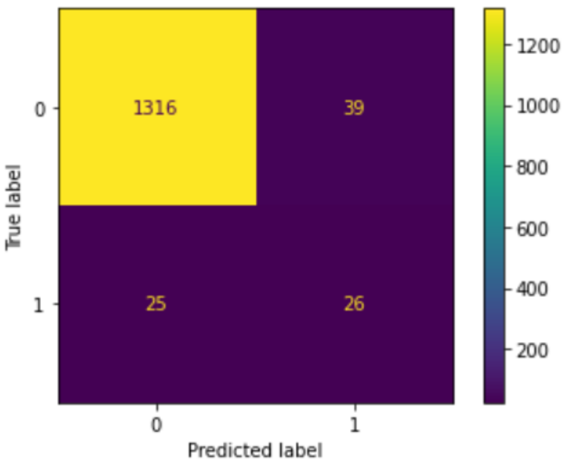
MODEL	ACCURACY
Logistic Regression	0.9602
Decision Tree	0.9545
Gaussian Naives Bayes	0.0718
Random Forest Classifier	0.9666
Boosting Classifier	0.9701
KNN Classifier	0.9609
Bagging Classifier	0.8962

CONFUSION MATRIX

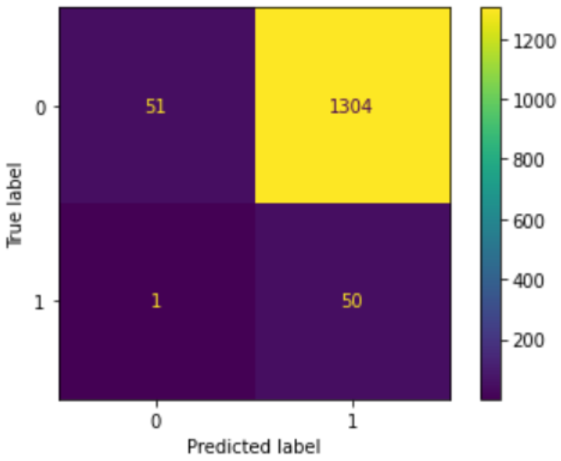
Logistic Regression



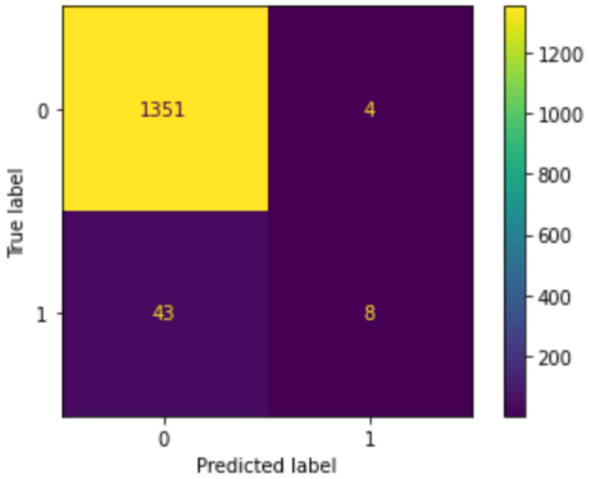
Decision Tree



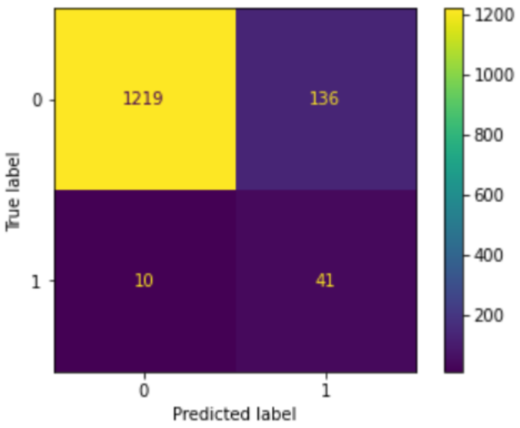
Gaussian



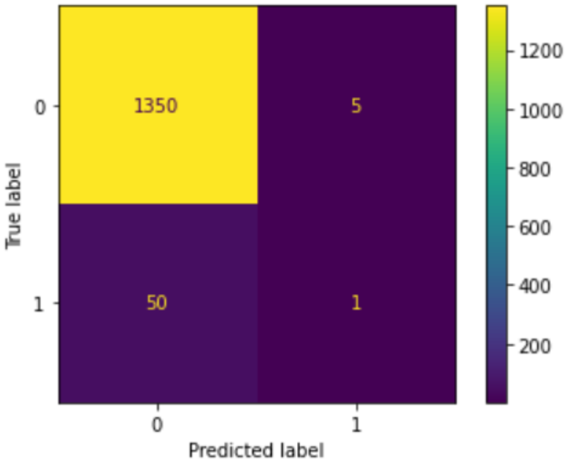
Random Forest



Bagging



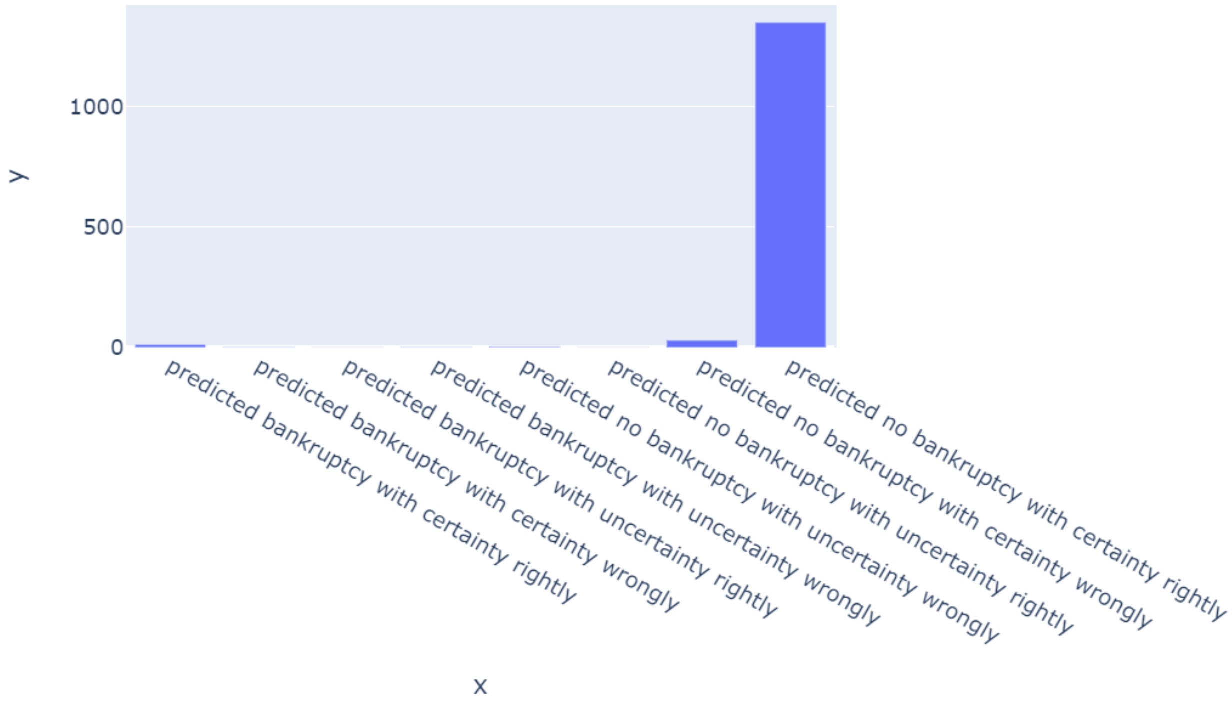
KNN

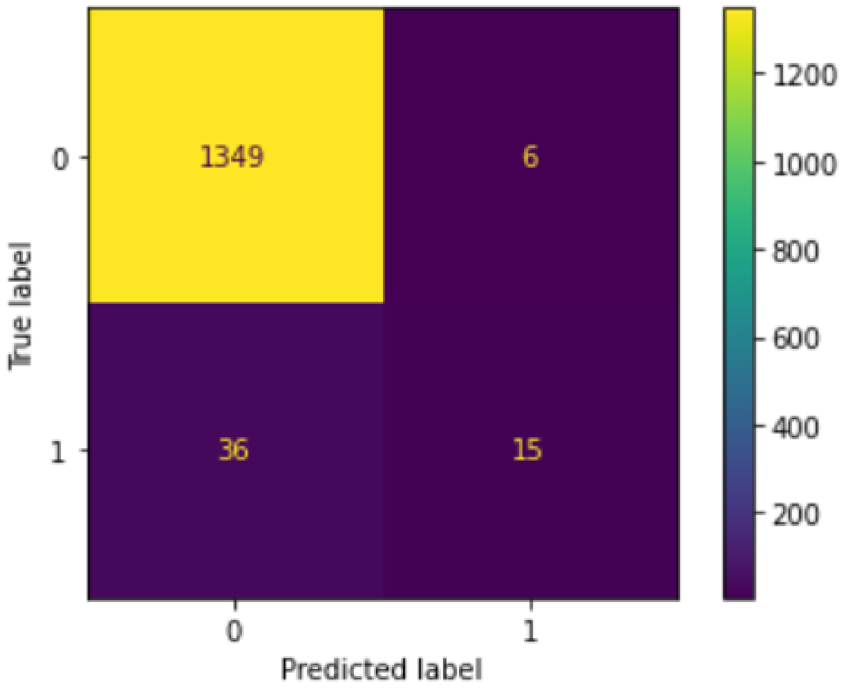


We decided to reject all of this model because of their result weren't as good as the ones of the boosting tree.

FAVORITE MODEL : BOOSTING

- Boosting Tree Classifier achieves a good result but it misses more than half bankruptcies.





- Our dataset has some well-defined bankrupted companies, it also has companies in good health and a major part of the bankrupted companies being close from not bankrupted companies.

BOOSTING TREE : PARAMETER TUNIG

- To assess during our cross validations, we use balanced accuracy instead of accuracy hence the lower value that you may see.
- We first split the dataset in two parts (one for cross validation and one for a final test).
- We have chosen kfold cross validation in order to get a better assessment of a model and its parameters. We will do 3 repetition for each parameter couple. 2 seemed not very robust and we found that 4 repetition was too much, 3 seemed to be a great compromise between ETA and Evaluation of the model. We've chosen 5 split to keep the classic ratio of 80% for training and 20% for testing.

K-FOLD CROSS VALIDATION

TEST 1

```
boostingParams={
    'n_estimators':[50,500],#1000 very slow
    'learning_rate': [0.05,0.1,0.2],
    'max_depth':[3,7,10],
    'max_features':['sqrt', 'log2'],# N feature very slow
    'min_samples_split':[0.2,0.4],
    'min_samples_leaf':[0.1,0.2]
}
```

	learning_rate	max_depth	max_features	min_samples_leaf	min_samples_split	n_estimators	test_score
97	0.20	3	sqrt	0.1	0.2	500	0.599853
115	0.20	7	sqrt	0.1	0.4	500	0.598694
131	0.20	10	sqrt	0.1	0.4	500	0.598578
99	0.20	3	sqrt	0.1	0.4	500	0.590665

TEST 2

```
boostingParams={
    'n_estimators':[400,600],#1000 very slow
    'learning_rate': [0.15,0.25],
    'max_depth':[3,8],
    'max_features':['sqrt', 'log2'],# N feature very slow
    'min_samples_split':[0.1,0.5],
    'min_samples_leaf':[0.05,0.1]
}
```

	learning_rate	max_depth	max_features	min_samples_leaf	min_samples_split	n_estimators	test_score
33	0.25	3	sqrt	0.05	0.1	600	0.639765
51	0.25	8	sqrt	0.05	0.5	600	0.637649
50	0.25	8	sqrt	0.05	0.5	400	0.632925
1	0.15	3	sqrt	0.05	0.1	600	0.628403

TEST 3

```
boostingParams={
    'n_estimators':[400,600],#1000 very slow
    'learning_rate': [0.18,0.23],
    'max_depth':[3,8],
    'max_features':['sqrt'],# N feature very slow
    'min_samples_split':[0.1,0.5],
    'min_samples_leaf':[0.02,0.005]
}
```

	learning_rate	max_depth	max_features	min_samples_leaf	min_samples_split	n_estimators	test_score
29	0.23	8	sqrt	0.005	0.1	600	0.678578
13	0.18	8	sqrt	0.005	0.1	600	0.678549
21	0.23	3	sqrt	0.005	0.1	600	0.675012
15	0.18	8	sqrt	0.005	0.5	600	0.668056

K-FOLD CROSS VALIDATION

TEST 4

```
boostingParams={
    'n_estimators':[700,1000],
    'learning_rate': [0.18,0.23],
    'max_depth':[3,8],
    'max_features':['sqrt'],# N feature very slow
    'min_samples_leaf':[0.02,0.005,0.0025]
}
```

RESULTS

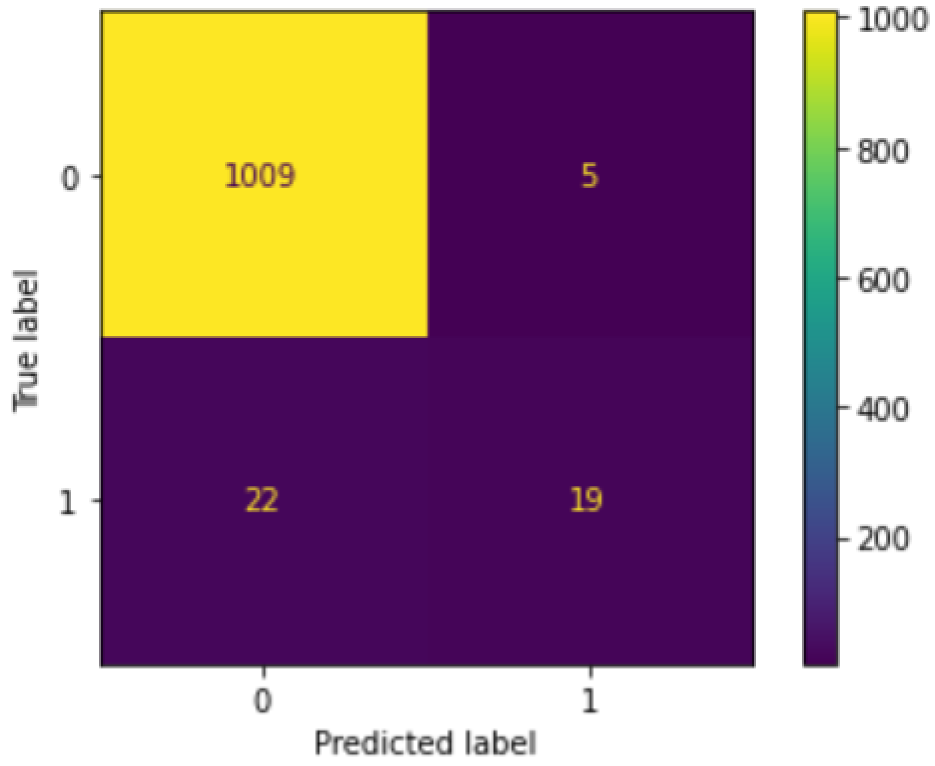
	learning_rate	max_depth	max_features	min_samples_leaf	n_estimators	test_score
17	0.23	3	sqrt	0.0025	1000	0.689795
15	0.23	3	sqrt	0.0050	1000	0.689157
5	0.18	3	sqrt	0.0025	1000	0.683447
3	0.18	3	sqrt	0.0050	1000	0.682868

ANALYSE

This time we haven't gained accuracy, it is time to stop.

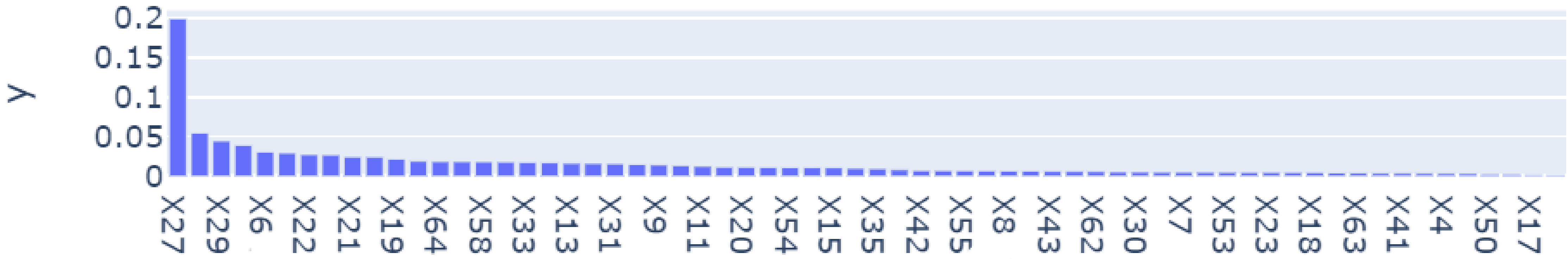
BOOSTING

- Finally, with our tuning parameters we test our model and we obtain a very good model with an accuray = 0.9744.



VARIABLES

- As we can see, X27 maximize the variable of the model



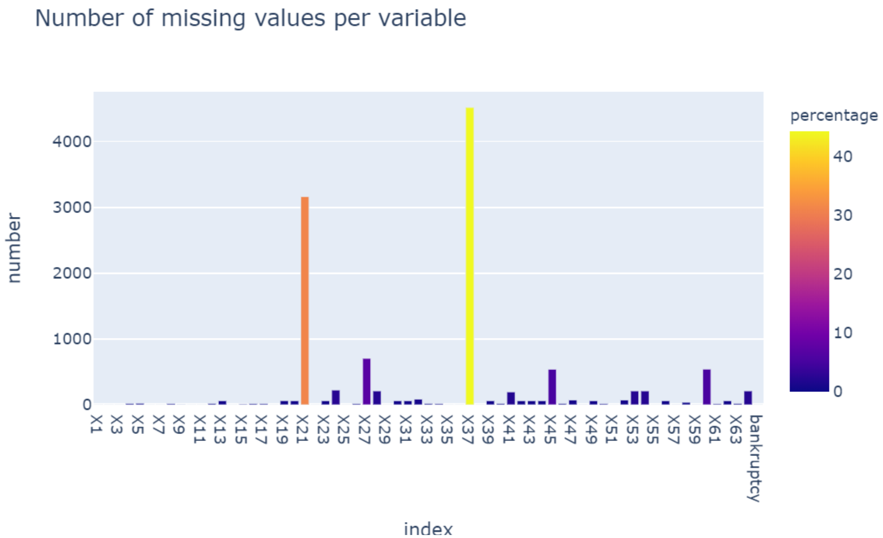
X27 = profit on operating activites / financial expenses
X29 = logarithm of total assets
X6 = retained earnings / total assets

5 DATASETS

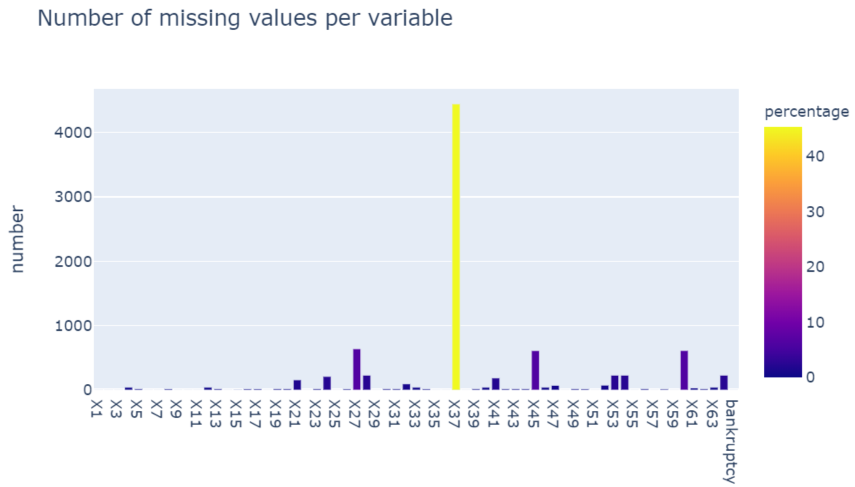
- For each data set, we do the same analyse so :
  - Visualize the number of missing values per variable
  - Visualize the mean rate and the medium rate of change
  - Make a cross validation to tune the parameters
  - Print the accuracy
  - Visualize the best represented variables
  - Print the confusion matrix

MISSING VALUES

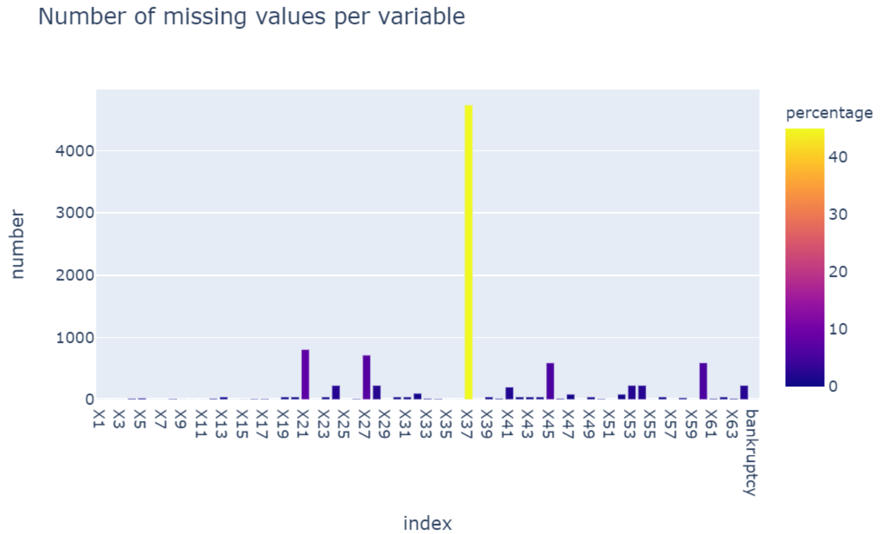
2nd year



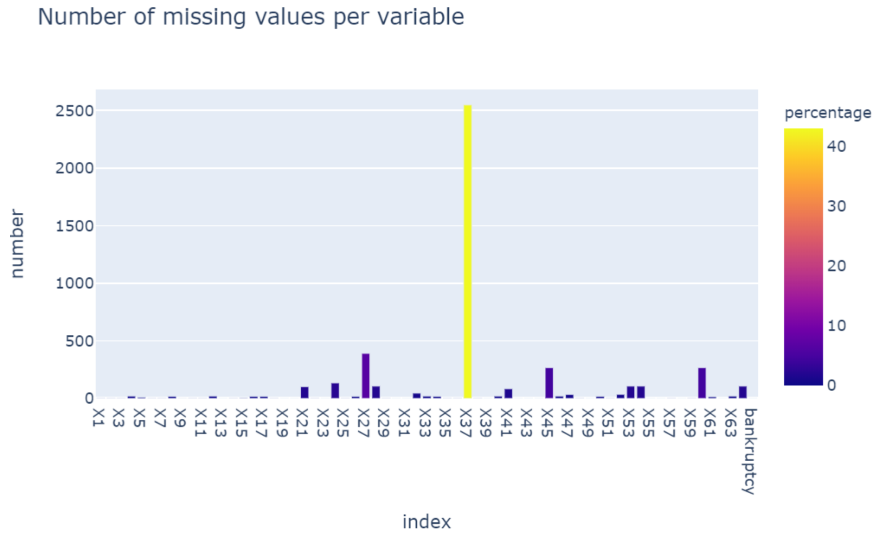
4nd year



3nd year

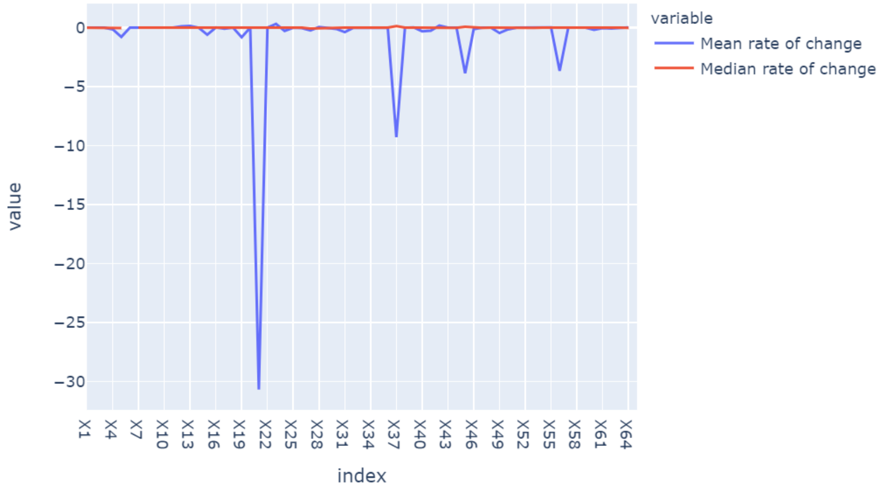


5nd year

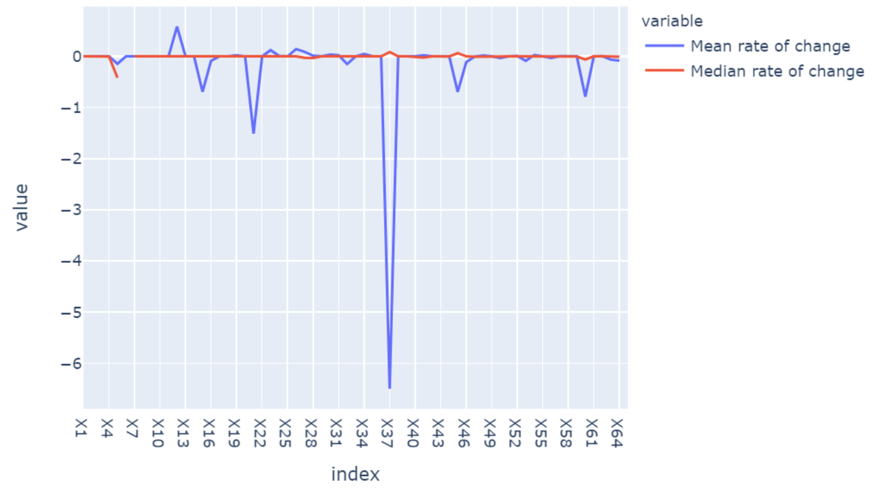


MEAN AND MEDIAN RATE OF CHANGE

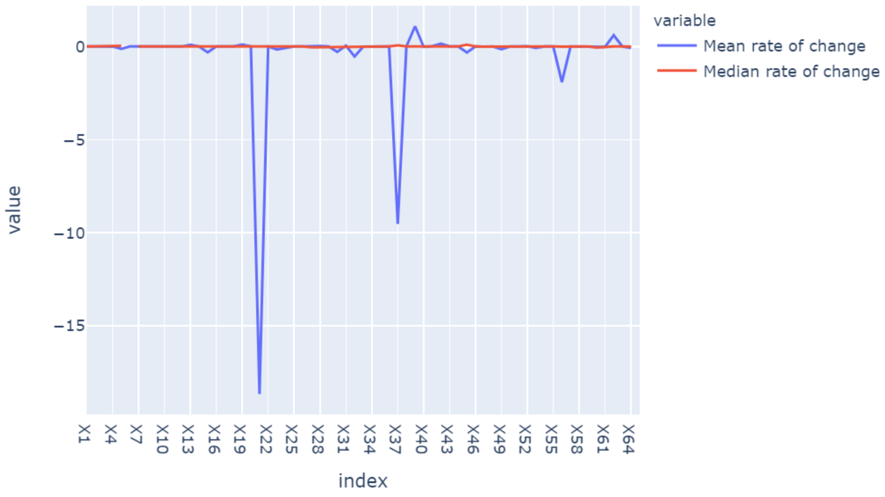
2nd year



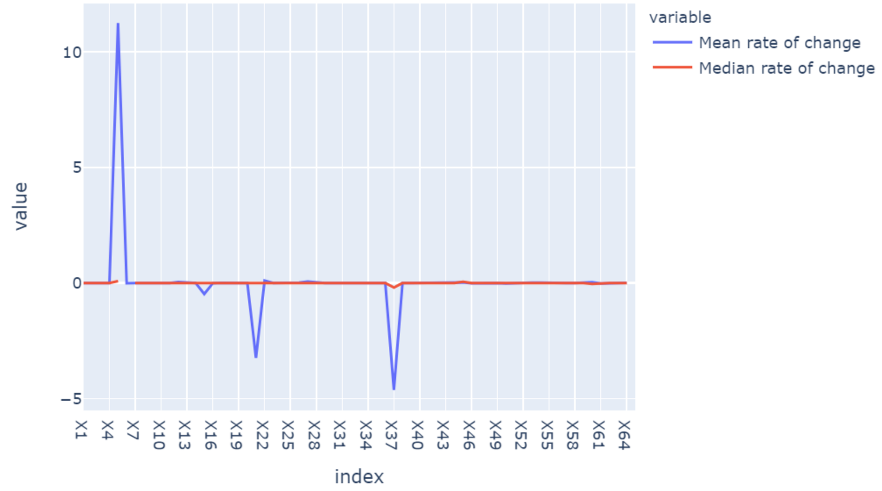
4nd year



3nd year



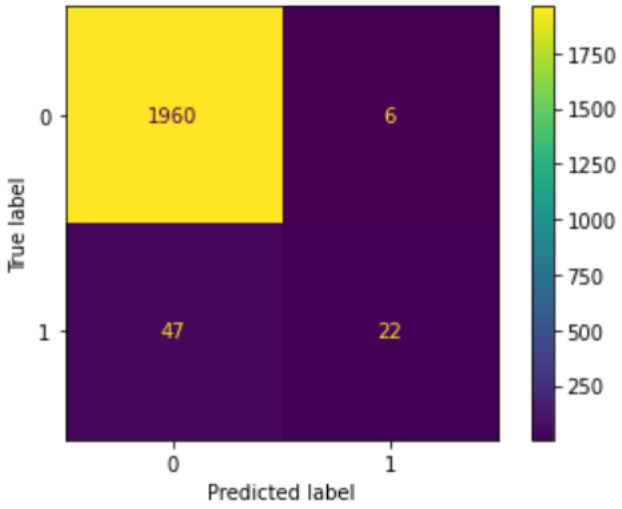
5nd year



ACCURACY / CONFUSION MATRIX

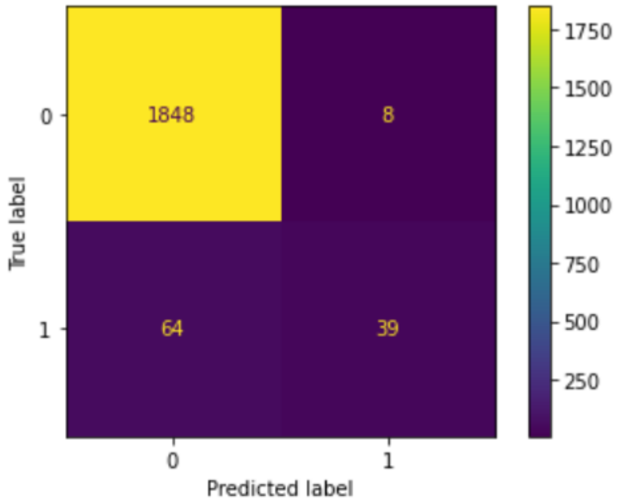
2nd year

Accuracy = 0,9740



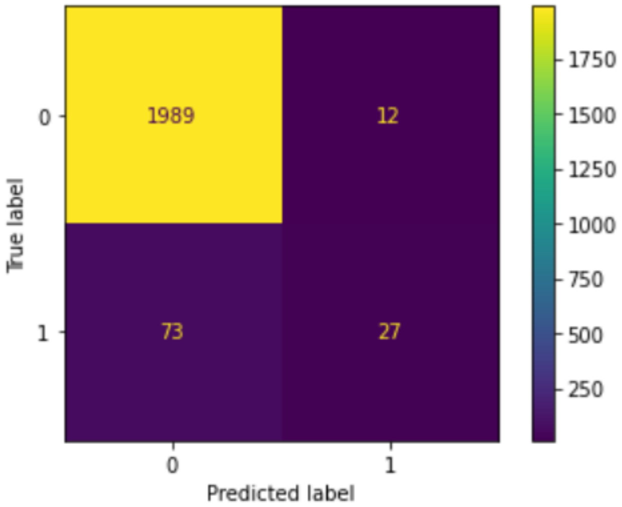
4nd year

Accuracy = 0,9632



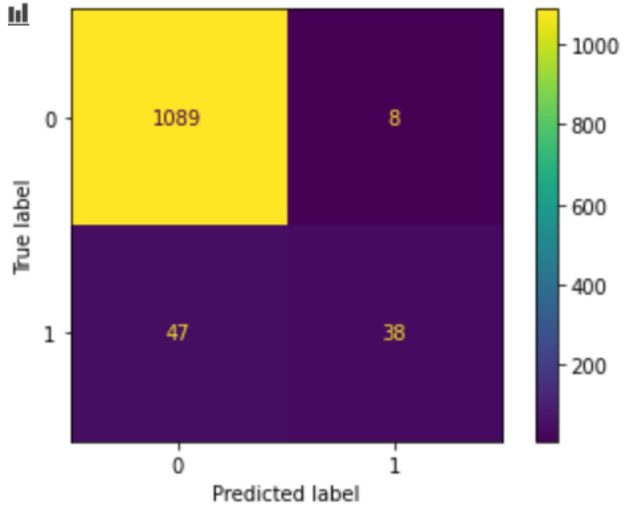
3nd year

Accuracy = 0,9595

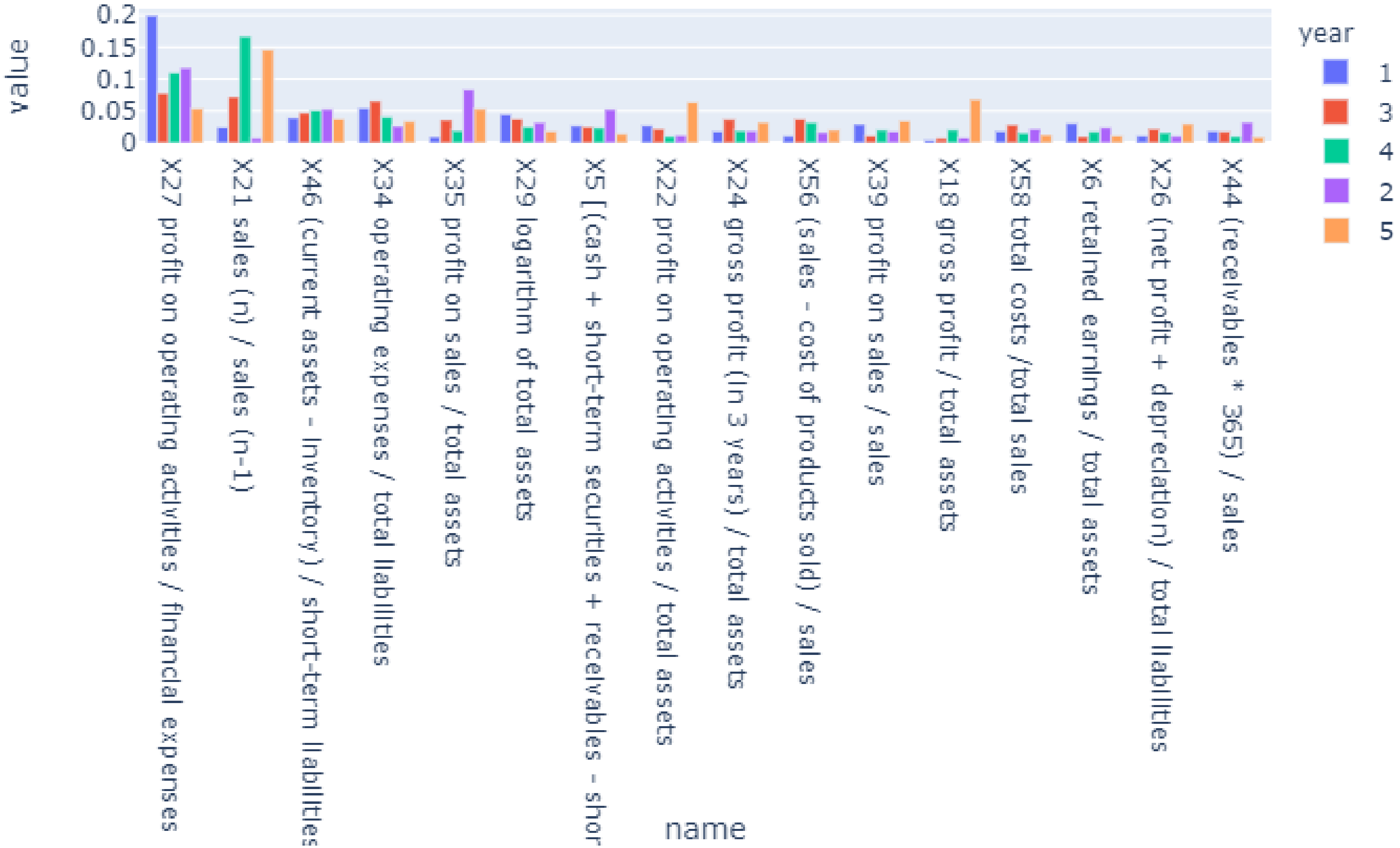


5nd year

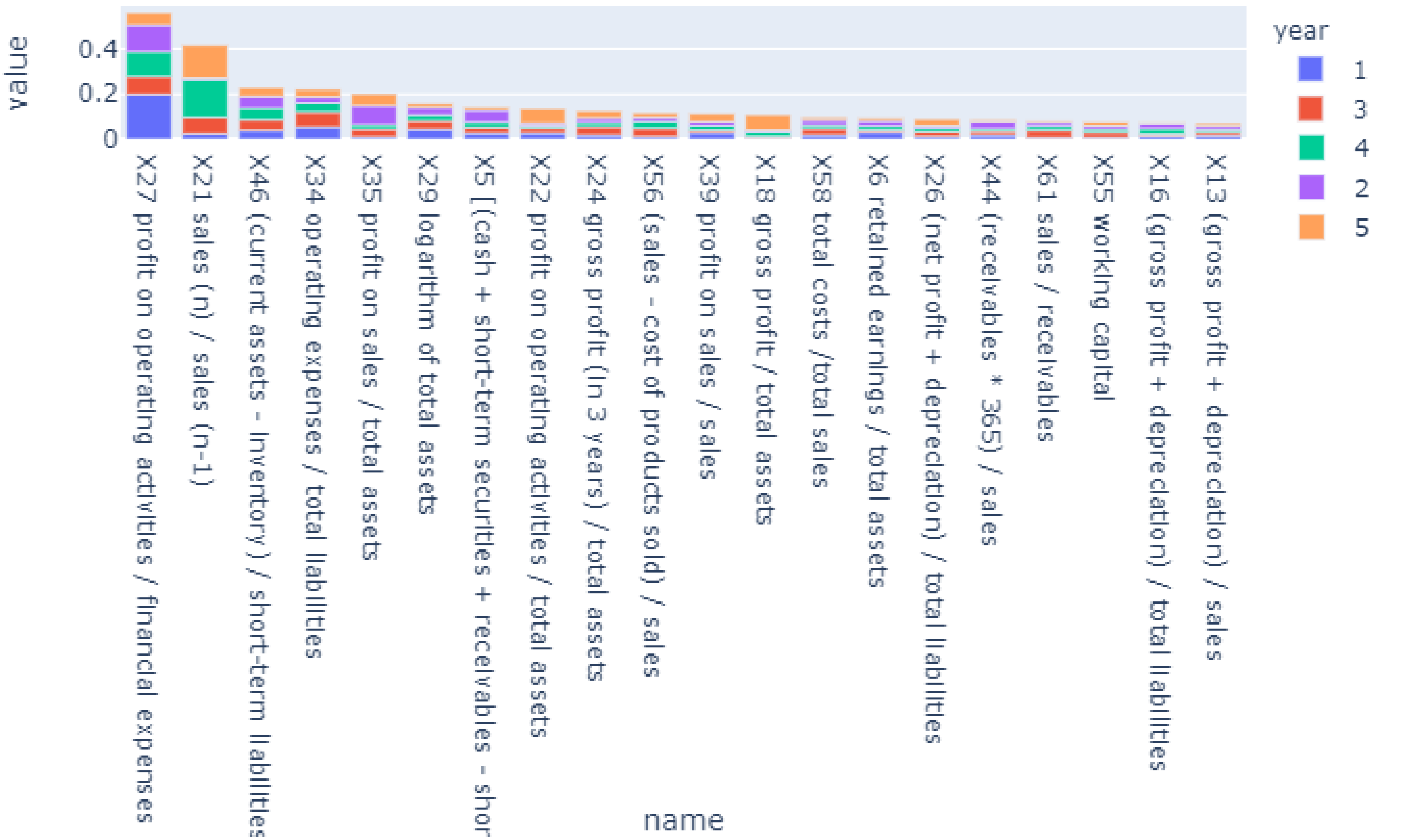
Accuracy = 0,9535



VARIABLES IMPORTANCE



VARIABLES IMPORTANCE



CONCLUSION

- This dataset was a challenging dataset
- First there was a lot missing data and it took us a lot of time before we could make tests with the integrality of the dataset
- Then our data was fragmented in several files depending on the year it was collected, we decided to keep this fragmentation and make 5 analysis (in fact one analysis generalized to all the others datasets in a later time)
- This approach was the right one as at the end we can see some differences on the variable importance between the different datasets. We have a great overall accuracy. But as there isn't that much bankrupted companies (thank you god) it is difficult for the model to catch up all the important points.
- In the website we do not ask enough variables to user which leads to extremely bad results
- Here is a list of things we could have done (better):
    - Duplicate the bankrupted data and add some noise add weight to this small class
    - Focus less on hyperparameters optimization it took us a lot of both computational and personal time to get a moderate amelioration
    - Remove some variables, we're used to remove insignificant variables in regressions but not in classification. This may come from the fact that our tree balances himself alone by choosing were he's cutting and that this step unnecessary.

Thank you for your attention Alexandre and Julie