Louis ATALLAH

Polish Companies Bankruptcy

The context

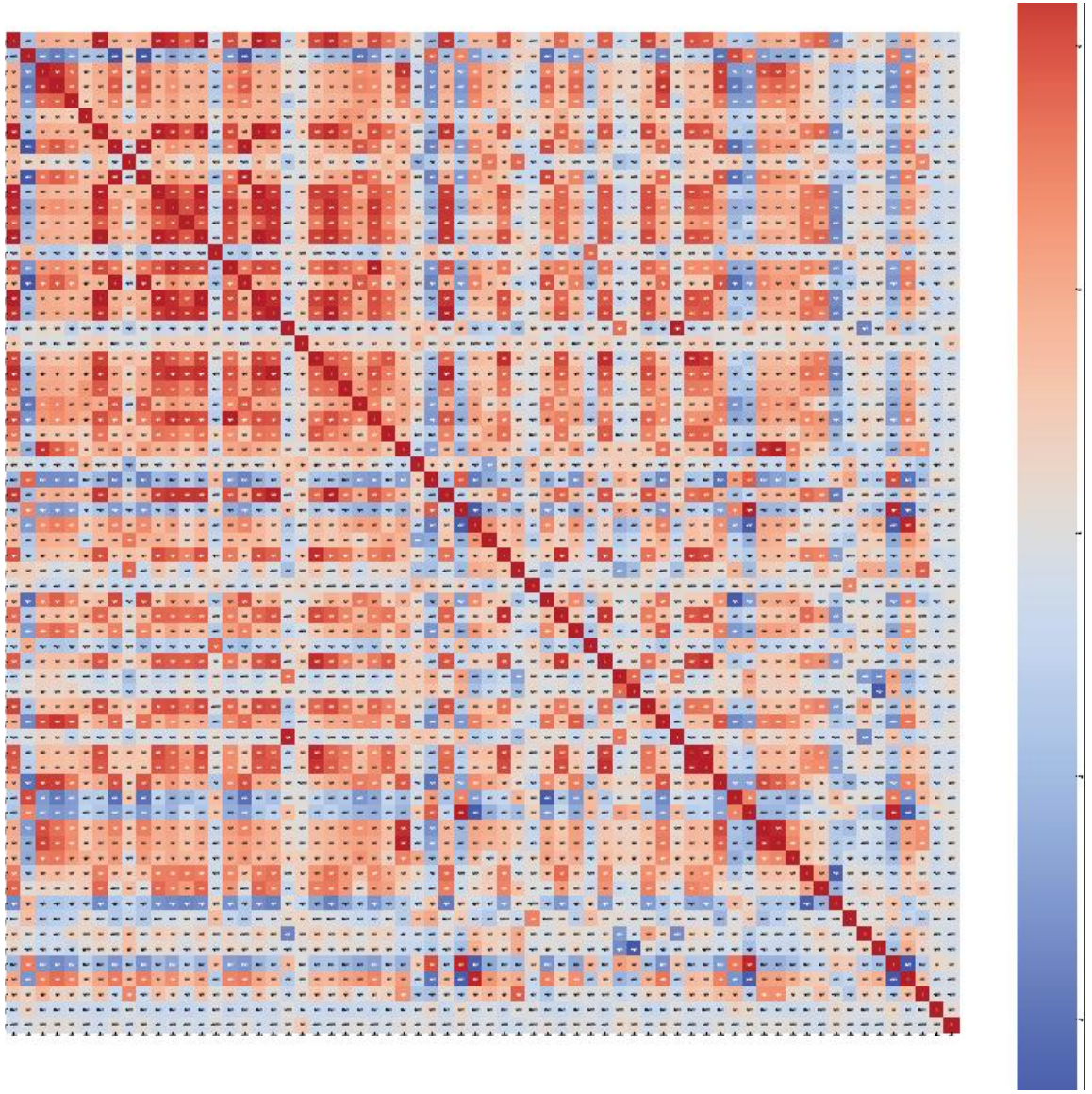
- *Classification : We need to predict whether the company will go to bankruptcy (simple classification)*
- *Dataset of 64 features*
- *Collected data from 2007 to 2013 on 43405 Companies*
- *Target named « class » : 1 for bankruptcy – 0 for not*

Data preprocessing

- *The dataset is composed by 5 .arff files, thus I needed to convert them into csv files, then merge them together.*
- *I have added a feature about the year to the corresponding initial files. It enables to not loose the tracabilitty of the year by using all in one.*
- *The original files contain some '?' that I had to replace by 0, I did for the NA values as well.*
- *I have converted the object type features into float type in order to manipulate them.*
- *I have scaled my dataset because some values are greater than others.*
- *No constant nor quasi constant features.*

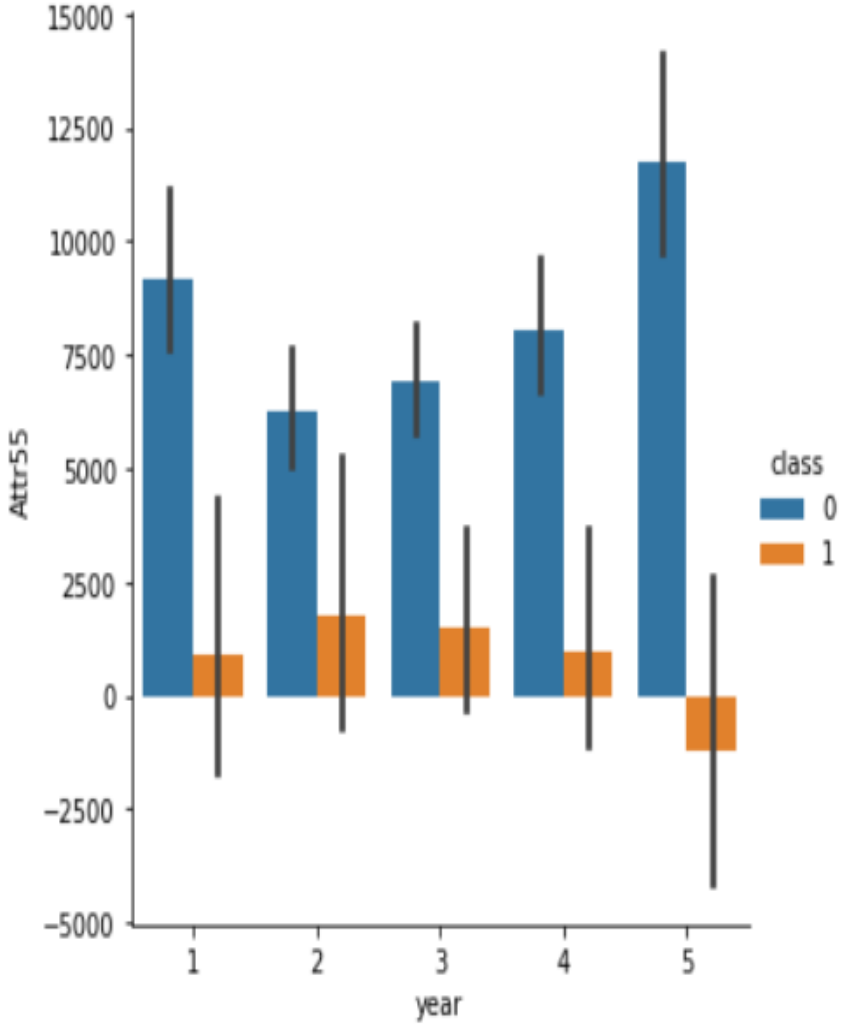
Data visualization

- *I have observed on the heatmap that many variables are correlated to each other (deep blue or red).*
- *In order to optimize my models, I dropped the correlated variables (32 unnecessary features)*



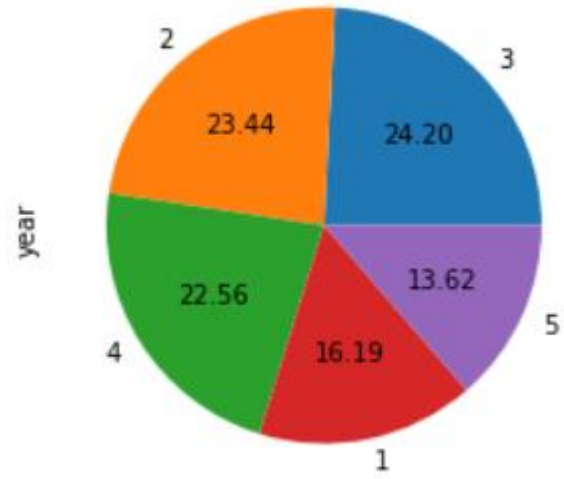
Data visualization

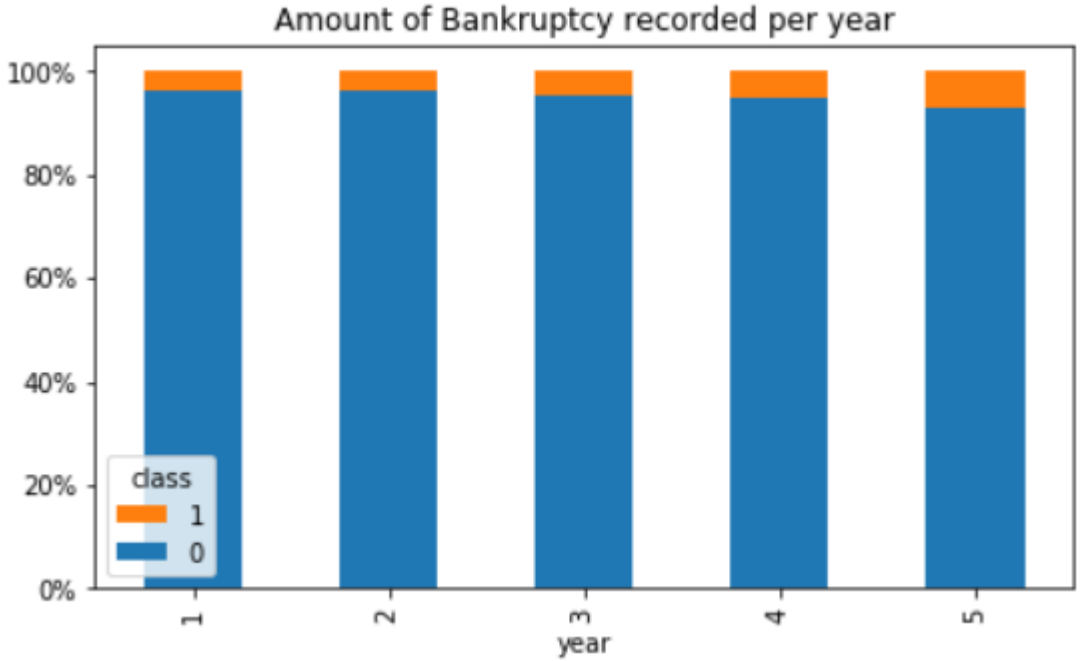
- *I have displayed some very interesting graphs to show the relationship with one of the attributes. The beside one is about the bankruptcy according to the working capital (Attr55) through each year. The working capital is clearly much lower for the bankrupt companies. Moreover, as we can see, for the second year, which represents 2008, we observe generally a decrease of the working capital for the polish companies. It can be explained by the financial crisis of the same year.*
- *Indeed, an average of a 31% loss for the polish companies working capital in 2008. We can remark as well that companies which went to bankruptcy in year 2 (2008) had a greater working capital than bankrupt companies from other years. We can infer the economic and financial context at this time. The contrary to year 5 when the majority of bankrupt companies had a negative working capital which means it was "easier" to avoid the bankruptcy. It can also be explained by the well being of companies that year with the highest working capital of the sample.*



Data visualization

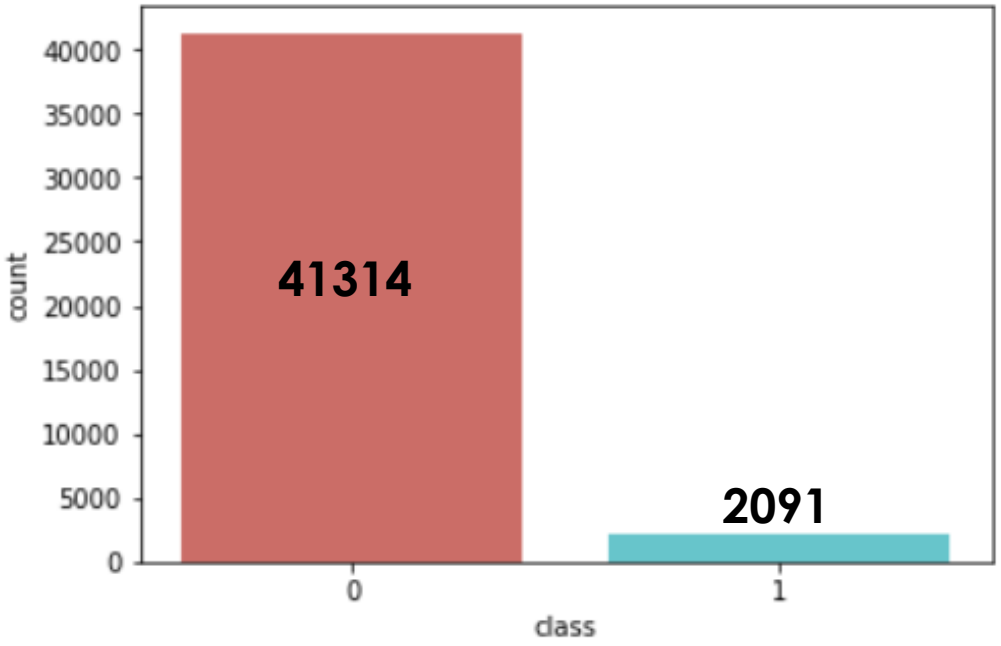
- *We can see on the pie chart besides, that the data through the years is well balanced. As it is shown by the percentages displayed, the number of companies per year are quite the same. There is no big difference.*
- *The lower graph is about the bankruptcy rate per year. It is almost the same for each year.*
- *By analyzing these two graphs, we can conclude that there is no problem to merge the five datasets in one because the rates per year are equivalent. Hence it will not unbalance the study.*





Data visualization

- *On the right, I have plotted the total number of companies which have not bankrupt and those which have.*
- *The number of no bankruptcy is very highly disproportionate, contrasting with the tiny part of the bankrupt companies.*
- *This characteristic leads us to remark that the classes are unbalanced.*
- *It will influence the choice of the metric.*



Fitting models

- *Once the data scaled, I split the dataset into two parts : the attributes and the target.*
- *Then I split the two parts in two to separate the train set and the test set with a repartion of 70% of the data for training and the other 30% for the testing set.*
- *Once done, I can start fitting some models.*
- *We have to bear in mind that the classes are unbalanced so we could not rely to the accuracy metric because if we do so we can have a model predicting full 0 and having an accuracy of 95%. We will see some examples of most interesting models.*
- *The accuarcy is not reliable, hence we have to look at other metrics such as the AUC or the recall (True positive rate).*

Logistic Regression

	precision	recall	f1-score	support
0	0.98	0.67	0.79	13650
1	0.09	0.66	0.16	674
accuracy			0.67	14324
macro avg	0.53	0.66	0.47	14324
weighted avg	0.93	0.67	0.76	14324

- *I used the class_weight parameter in the logistic regression function by setting it to "balanced". This will enable the model to take into account the unbalanced classes between 0 and 1 in the dataset.*
- *Classification report gives the accuracy and the recall, the metrics we are looking for. As we can see, it gives a very good recall of 66% which means this model predicts 66% of the bankruptcy. On the other hand, we can see that the model provides a good accuracy as well. It means that the model is credible for both cases of prediction.*

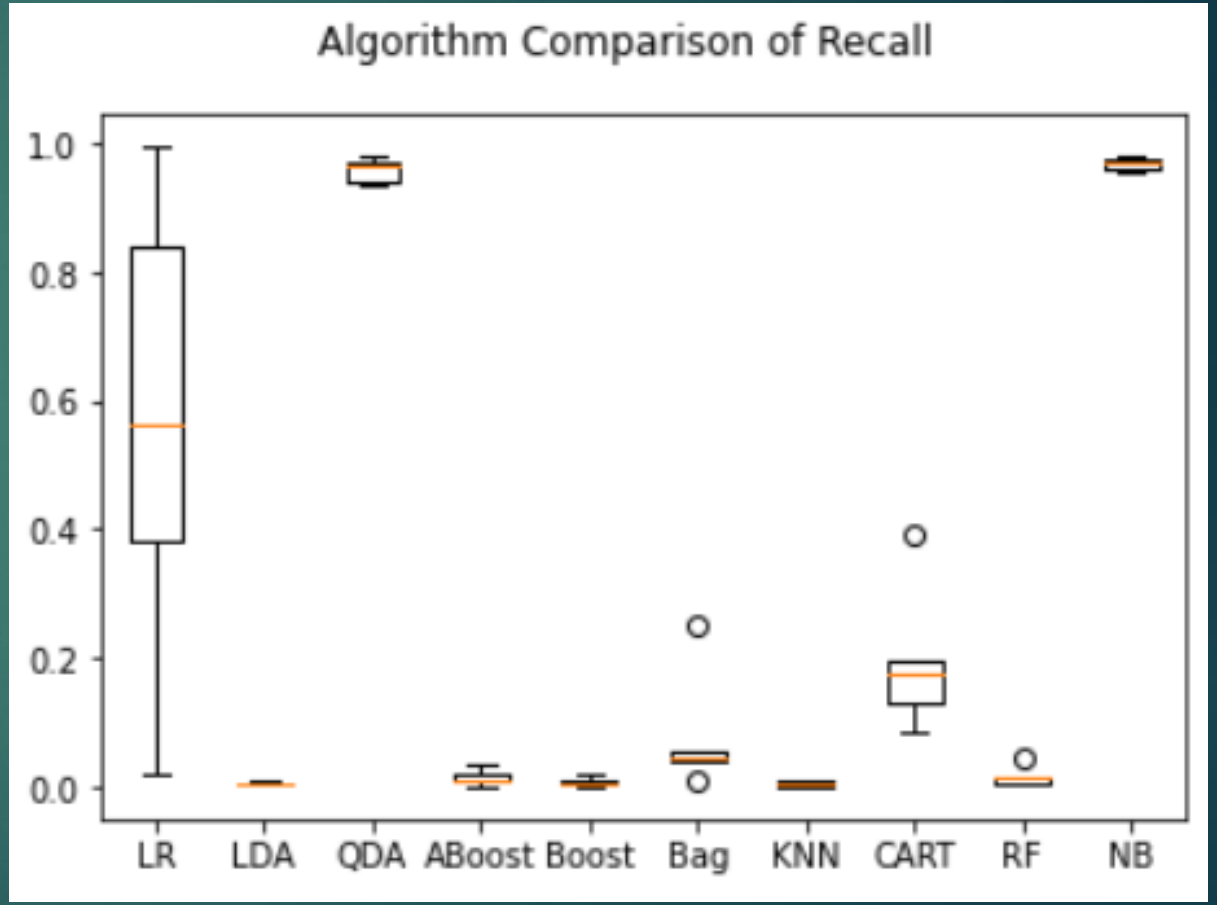
Gaussian Naive Bayes

	precision	recall	f1-score	support
0	0.94	0.23	0.36	13650
1	0.04	0.72	0.08	674
accuracy			0.25	14324
macro avg	0.49	0.47	0.22	14324
weighted avg	0.90	0.25	0.35	14324

- *The Naïve Bayes model is the total opposite of the precedent example. Indeed, it computes a strong recall which means it predicts almost every bankruptcy. The main drawback is the poor accuracy, the number of the false positives are too much important. This model predicts essentially 1. It can be misleading because it will announce the bankruptcy for a lot of companies which actually are not concerned.*

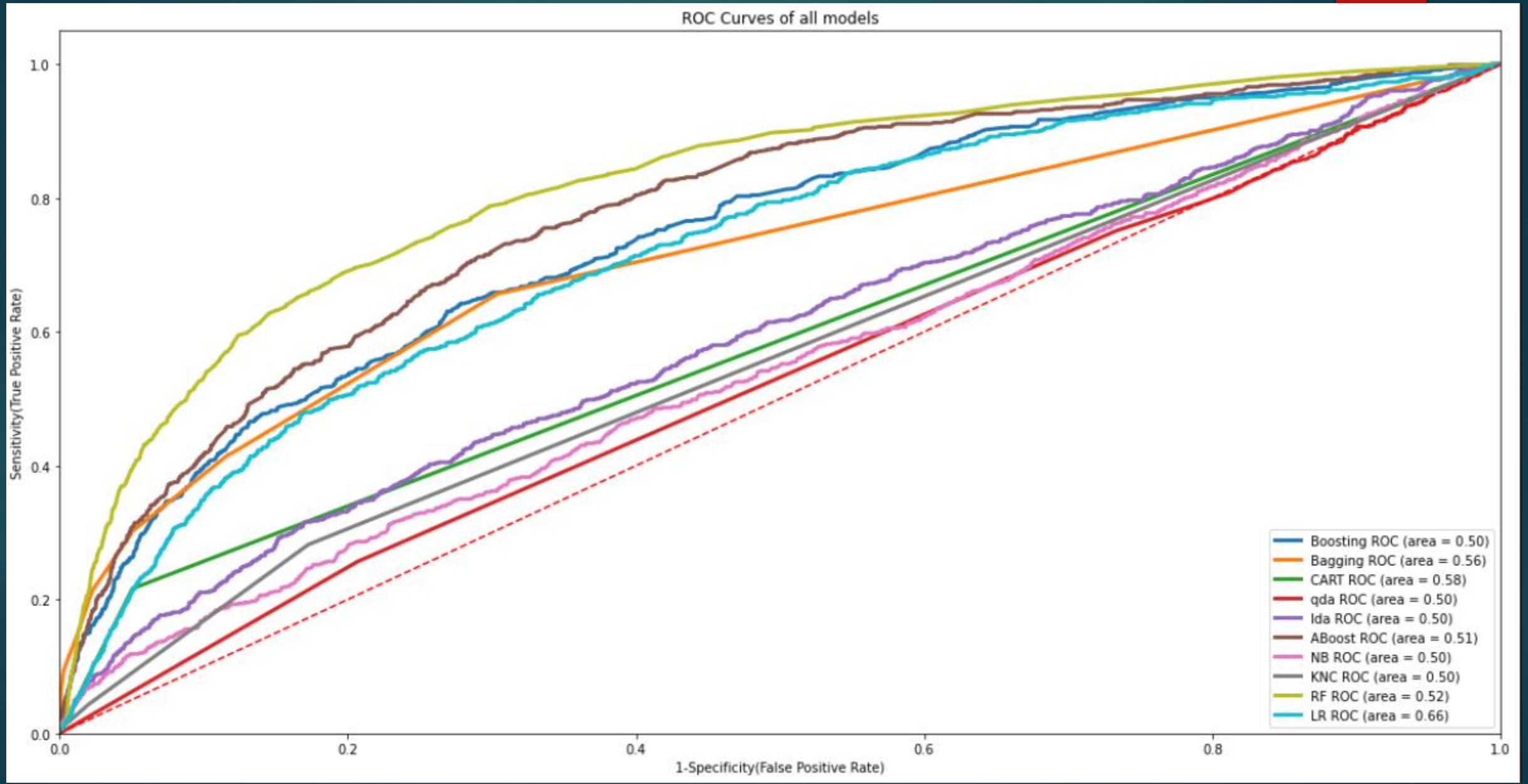
Models & Recall

- *Beside we can see the comparison of the recall of each model. The QDA and the Gaussian Naïve Bayes are similar, they have the best recall.*
- *We remark the third position of the Decision Tree Classifier (CART).*
- *Most of other models are very bad at predicting the 1 despite their good accuracy.*



ROC Curves

- *I have implemented a graph displaying every ROC curve of each model (on the next slide).*
- *The ROC Curve shows how rapidly a model is able to reach the top of the plot, the faster it does the better it is. It evaluates the recall according to the specificity (False positive rate).*
- *By doing so, we observe that the best ROC is from the Random Forests model, followed closely by the AdaBoost's and the Boosting's.*
- *We also remark that the worst model from this metric are the ones which have the best recall, the QDA and the Naive Bayes. It is mainly explained by their high rate of false positive predictions.*


ROC Curves of all models
Sensitivity(True Positive Rate)
1-Specificity(False Positive Rate)
0.0
0.2
0.4
0.6
0.8
1.0
Boosting ROC (area = 0.50)
Bagging ROC (area = 0.56)
CART ROC (area = 0.58)
qda ROC (area = 0.50)
lda ROC (area = 0.50)
ABoost ROC (area = 0.51)
NB ROC (area = 0.50)
KNC ROC (area = 0.50)
RF ROC (area = 0.52)
LR ROC (area = 0.66)

My choice for the API

- *To design my API, I have chosen the logistic regression model.*
- *My choice has been based on a good compromise between the AUC (66%) and the recall (68%).*
- *I have set up the model with the most important features in order to use them in the API.*
- *The 7 selected features are the following :*

- *(gross profit + depreciation) / total liabilities (Attr16)*
- *(equity - share capital) / total assets (Attr25)*
- *operating expenses / short-term liabilities (Attr33)*
- *(current assets - inventory - receivables) / short-term liabilities (Attr40)*
- *(current assets - inventory) / short-term liabilities (Attr46)*
- *sales / short-term liabilities (Attr63)*
- *Year of the study (2007 to 2012)*

	precision	recall	f1-score	support
0	0.97	0.59	0.74	13650
1	0.08	0.68	0.14	674
accuracy			0.60	14324
macro avg	0.53	0.64	0.44	14324
weighted avg	0.93	0.60	0.71	14324