Firm’s default prediction with Machine Learning

Tesi Aljai - Aris Anagnostopoulos - Stefano Piersanti

Outline

- *Firm's default prediction*
- *Datasets and Target variable*
- *ML techniques*
- *Experimental results*
- *Conclusion*

Firms default prediction

The problem

- An important issue faced for a long time. But also an increasingly relevant problem
- The “Non performing loans” (NPL) are about 20% of the total for Italian companies in recent years
- An improvement in loan default prediction accuracy can lead to savings of tens of billions of euros

Firms default prediction

History and related works

- Statistical/ML methods
- Financial data/Credit data

Bankruptcy decision techniques	
Statistical techniques	**Machine learning (ML) techniques**
Linear Discriminant Analysis (LDA)	Artificial Neural Network (ANN)
Multi Discriminant Analysis (MDA)	Support Vector Machine (SVM)
Logistic Regression (LR)	Decision Tree (DT)
	Random Forest (RF)
	Other ML Techniques

Bankruptcy decision data	
Financial data	**Credit data**
Total asset	Type of loans
Liabilities	Overdraft
ROE	Credit card data
ROA	

Datasets

CREDIT DATASET

- CCR dataset: about 1000 Italian banks and financial companies that give credit to firms
- Over 800.000 Italian firms
- High Granularity (single borrower)
- Quarterly data

BALANCE SHEET DATASET

- A small set of basic balance sheet indicators for a subset of medium/large Italian firms (about 300,000 firms)
- Frequency: annual data

The dataset: main attributes

ID	Description	ID	Description
L1	Granted amount of loans	**B1**	Revenues
L2	Used amount of loans	**B2**	ROE
L3	Bank's classification of firm	**B3**	ROA
L4	Average amount of loan used	**B5**	Total turnover
L5	Overdraft	**B6**	Total assets
L6	Margins	**B7**	Financial charges/operating margin
L7	Past due (loans not returned after the deadline)	**B8**	EBITDA
L8	Amount of problematic loans		
L9	Amount of non-performing loans		
L10	Amount of loans protected by a collateral		
L11	Value of the protection		
L12	Amount of forborne credit		

Table 1. Main attributes for the loan (L) and the balance-sheet (B) datasets.

BASELINE

The most relevant features are Bank's classification and Past due for Loan dataset and ROE and ROA for Balance sheet dataset.

Target variable

TARGET VARIABLE: **Firm's Adjusted Default Status:**

- A firms classification for the whole banking system
- A supervisory definition based on quantitative criteria relating to deteriorated loans

	Firms data (until time T)	Default at time T	Default at time T+1	TARGET
Firm 1	**Credit data+Balance sheet data**	NO	NO	0
Firm 2	**Credit data+Balance sheet data**	NO	YES	1
Firm 3	**Credit data+Balance sheet data**	YES	X	NOT CONSIDERED

- good credit condition at time T → default status at time T+1 year
- the problem is a "binary classification problem"

ML techniques

Decision tree

- Non-linear pattern classification
- simplicity and possibility of interpreting the result

Ensemble methods

Ensemble: use multiple learning algorithms to obtain better predictive performance

- **Random forest**
- **AdaBoost**
- **Gradient boosting**
- **Bagging**

«Combination» of classifiers

Evaluation of results

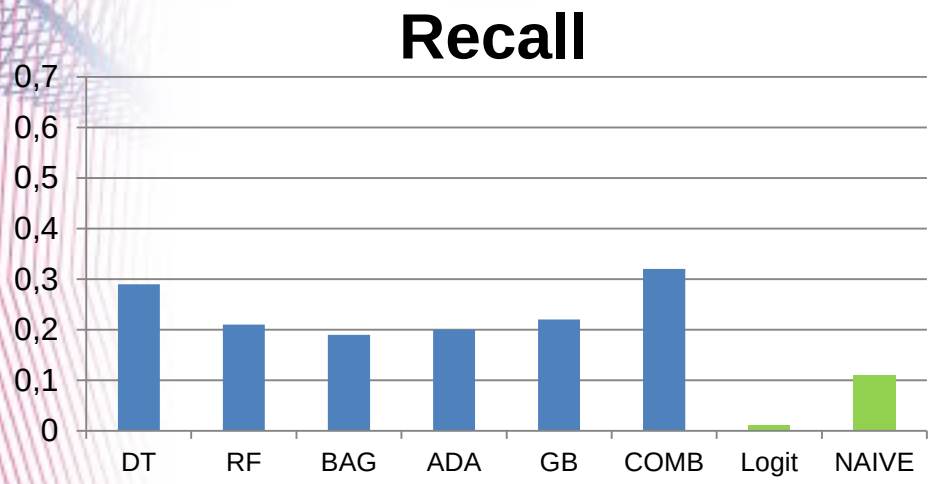
- **True Positive (TP)** *positive successful classification*
- **True Negative (TN)** *negative successful classification*
- **False Positive (FP)** *positive wrong classification, Type I error*
- **False Negative (FN)** *negative wrong classification, Type II error*

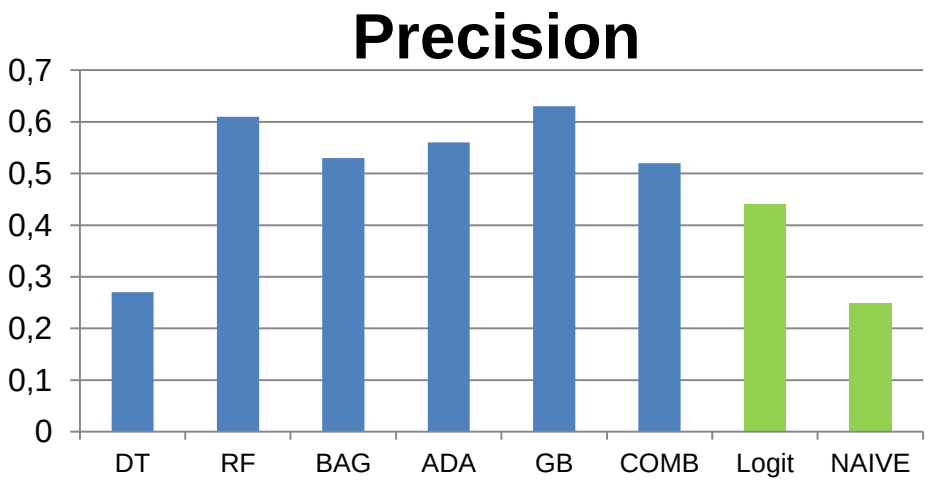
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Recall	$\frac{TP}{TP + FN}$
Precision	$\frac{TP}{TP + FP}$
F1 score	$2 * \frac{Precision * Recall}{Precision + Recall}$
Type I err	$\frac{FN}{TP + FN}$
Type II err	$\frac{FP}{TN + FP}$

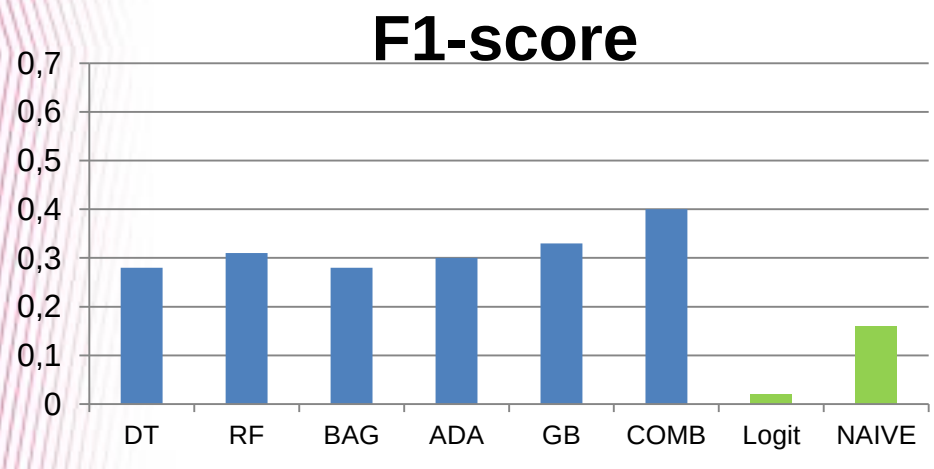
TARGET	PREVISION	
0	1	**FP**
1	1	**TP**
0	0	**TN**
1	0	**FN**

Experimental results

Adjusted Default prediction







DT = Decision tree

RF=Random Forest

BAG=Bagging

ADAB=ADaBoost

GBOOST=Gradient Boosting

NC=Naive classification

Logit=Logistic regression

ML TECHNIQUES

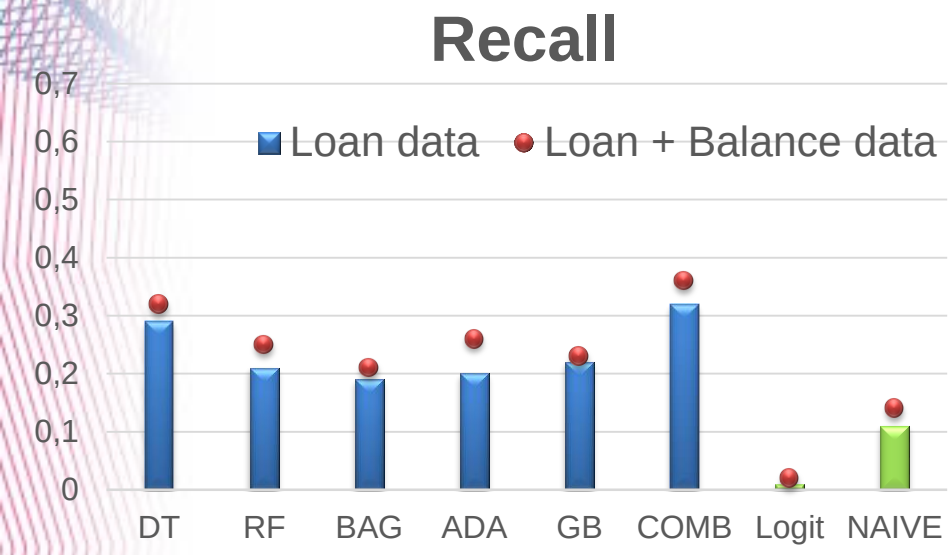
BASELINE

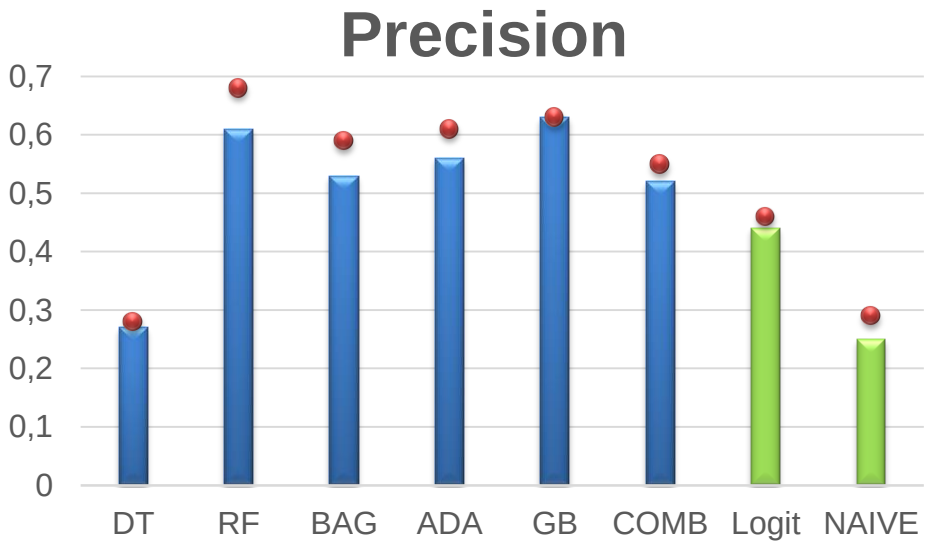
LOAN DATA - *IMBALANCED TRAINING SET*

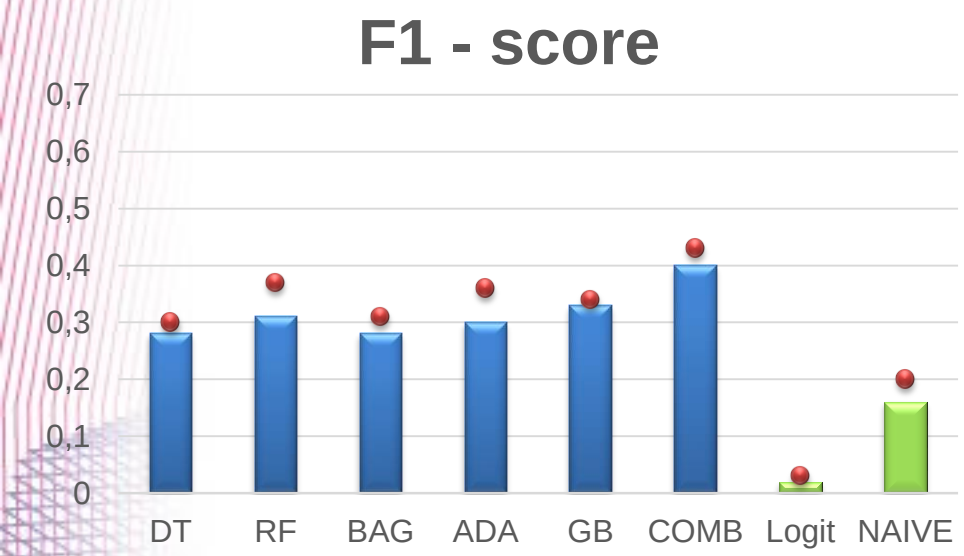
- Higher Precision – Lower Recall
- Combined method shows better performance for F1-score

Experimental results

Imbalanced training set



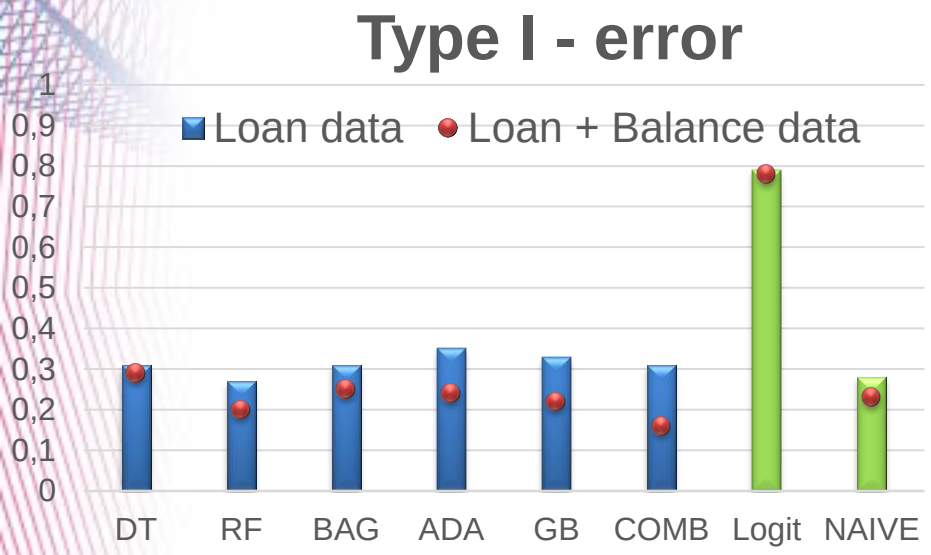


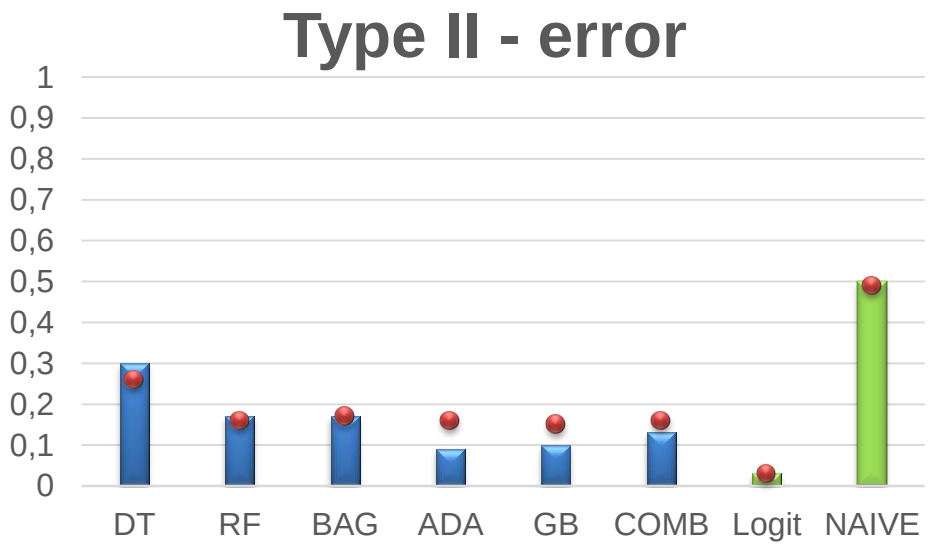


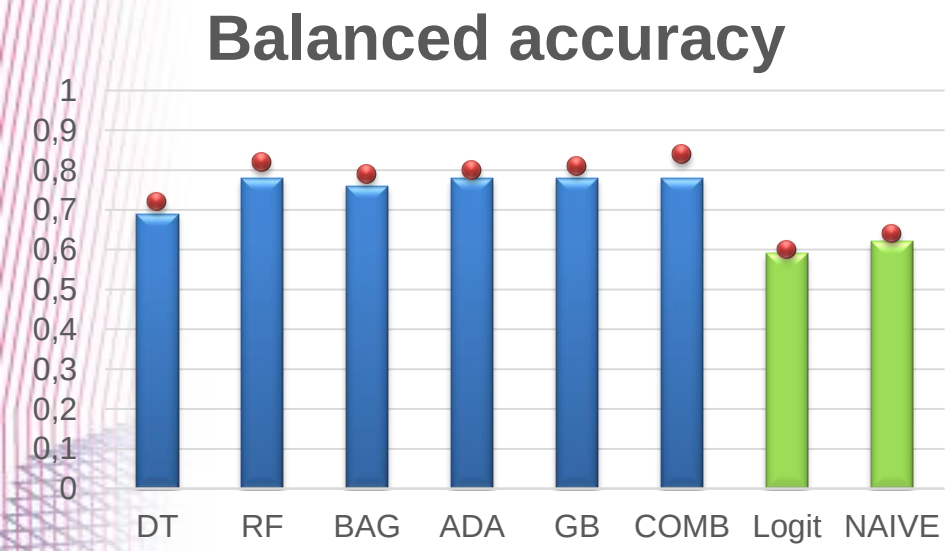
- Combined method shows better performance
- Use of balance sheet data slightly improves performance

Experimental results

Balanced Training set



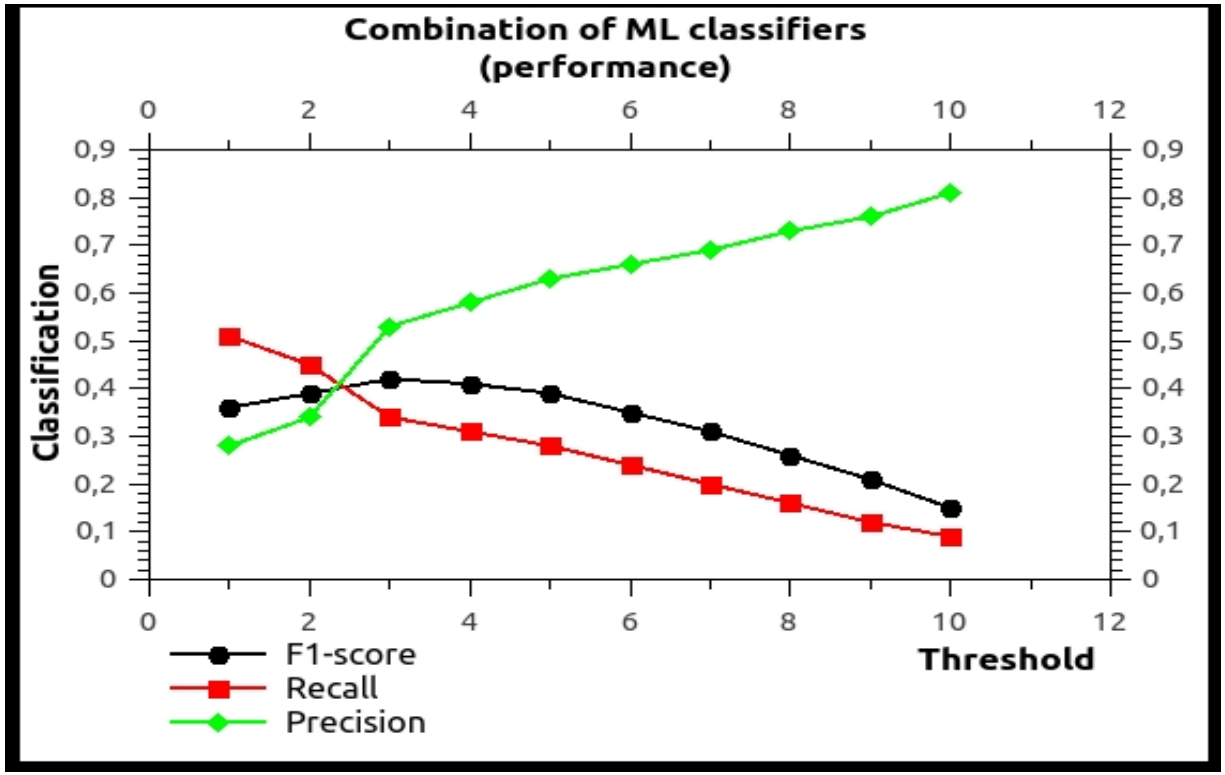




- **Higher Recall** – Lower Precision We try to minimize T1 and T2 errors
- Combined method shows the best performance for T1-err and T2-err (0.16 for both)
- Use of balance sheet data improves performances for all classifiers

Combination of classifiers

Adjusted Default Prediction



In order to improve the performance: "Combine" the previous techniques

- Increase the classification threshold → advantage in Precision
- However, we will identify less default firms, with a loss in Recall.

Preference for Precision

A practical application

Estimate of PD (Probability of Default): a practical application

- **"Baseline"** is a commonly used method for estimating the probability of default (PD): $\frac{New\ default\ loan\ at\ time\ T}{Total\ loans\ at\ time\ T-1}$
- **"COMB"** is the new "Combined" ML method
- **"Coarse segmentation"** has about 400 subsets while **"Fine segmentation"** over 10,000

Coarse segmentation	Baseline	COMB
Mean error	0.11	0.048
Var error	0.056	0.016
Superiority percentage	25.1%	45.6%

Fine segmentation	Baseline	COMB
Mean error	0.088	0.036
Var error	0.06	0.025
Superiority percentage	6.1%	19.5%

The experimental results show that **the new ML classification method works generally better.**

Summary and Conclusion

- *A **first attempt** to explore the Italian Central Credit Register dataset using ML techniques **in order to predict firms Adjusted Default Status.***
- *We combine credit data **with balance sheets data**, showing that we can improve further the accuracy of predictions.*
- *Moreover, **combining classifiers of different type** can lead to even better results.*

Next steps

Try to improve the forecast performance along two lines:

- Use also other ML techniques
- Use a greater number of balance sheet data

Try to extend at:

- Bankruptcy prediction with credit data

… The end…

Thank You

Target variable

BANKRUPTCY versus ADJUSTED DEFAULT STATUS

Firms Bankruptcy

- It is a juridical definition
- Connection between Bankruprcy and Adjusted default in our dataset

	Default	No Default	Total
Total	13200	290000	303200
No Bankruptcy	11040	283040	294080
Bankruptcy	2160	6960	9120
% Bankruptcy	16.4	2.4	

Analysis of results

Adjusted Default Prediction

We are working with a highly unbalanced dataset

- Balanced versus Imbalanced training set
- Recall versus Precision
- F1 score versus Balanced Accuracy

Experimental results

Adjusted Default prediction

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.25	0.11	0.16	0.89	0.04	0.54
MNB	0.95	0.05	0.09	0.95	0.02	0.52
LOG	0.44	0.01	0.02	0.99	0.01	0.50
GB	0.63	0.22	0.33	0.78	0.01	0.61
RF	0.61	0.21	0.31	0.79	0.01	0.60
DT	0.27	0.29	0.28	0.71	0.03	0.63
BAG	0.53	0.19	0.28	0.81	0.01	0.59
ADA	0.56	0.20	0.30	0.80	0.01	0.60
COMB	0.52	0.32	0.40	0.68	0.01	0.66

Table 3. Imbalanced training set; loan data. Higher values are better, except for Type-I and Type-II error.

LOAN DATA - ***IMBALANCED TRAINING SET***

- Higher Precision – Lower Recall
- The best classifier is Gradient Boosting
- F1 score reaches higher values and the combined method shows better performance