Classification Project

Nicola Szwaja
Piotr Droś

Goal of the project

1

- Analyse the data

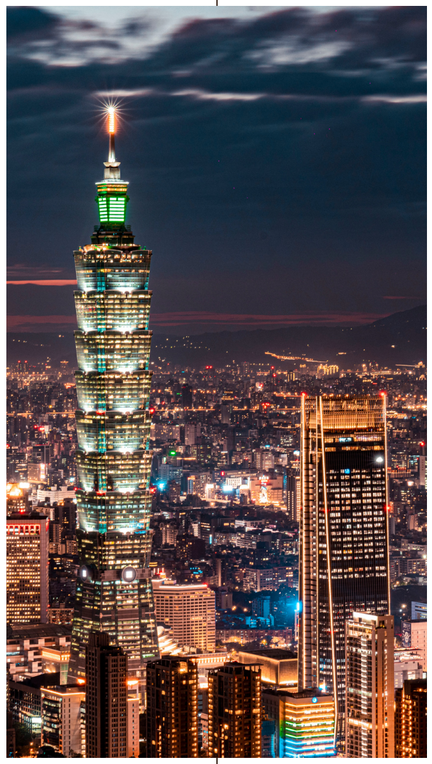
2

- Pick what the model will predict
- Process the data

3

- Create models
- Pick the best one

Dataset overview



The data were collected from the Taiwan Economic Journal for the years **1999 to 2009**. Company bankruptcy was defined based on the business regulations of the **Taiwan** Stock Exchange.

Dataset overview

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	...	Net Income to Total Assets	Total assets to GNP price	No-credit Interval	Gross Profit to Sales	Net Income to Stockholder's Equity	Liability to Equity	Degree of Financial Leverage (DFL)	Interest Coverage Ratio (Interest expense to EBIT)	Net Income Flag	Equity to Liability
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	...	0.716845	0.009219	0.622879	0.601453	0.827890	0.290202	0.026601	0.564050	1	0.016469
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556	...	0.795297	0.008323	0.623652	0.610237	0.839969	0.283846	0.264577	0.570175	1	0.020794
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	0.796403	0.808388	0.302035	...	0.774670	0.040003	0.623841	0.601449	0.836774	0.290189	0.026555	0.563706	1	0.016474
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	0.796967	0.808966	0.303350	...	0.739555	0.003252	0.622929	0.583538	0.834697	0.281721	0.026697	0.564663	1	0.023982
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	...	0.795016	0.003878	0.623521	0.598782	0.839973	0.278514	0.024752	0.575617	1	0.035490

Instances

6819

Features

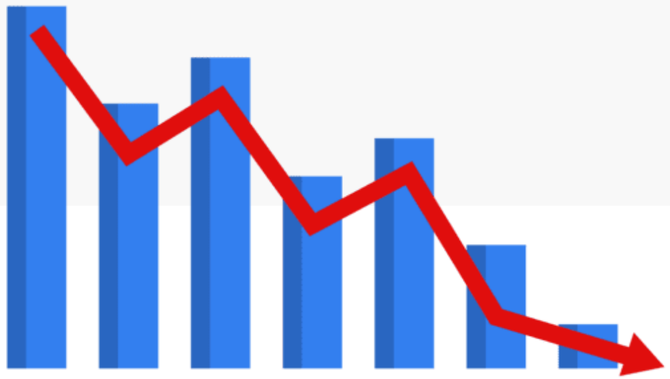
96

...

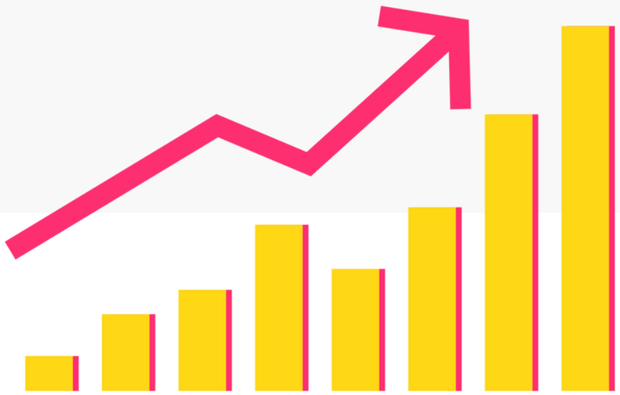
What the model will predict?

Definition of classes in the project

Class 1: "Bankrupt"
Model predicts the company is at risk of bankruptcy.

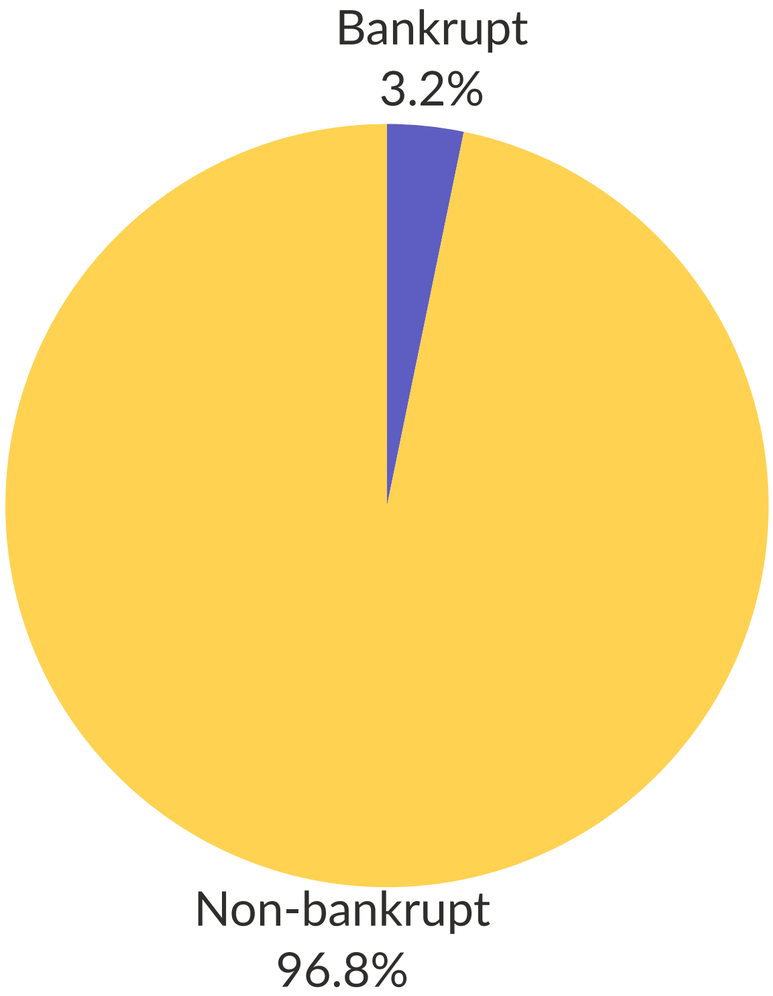


Class 0: "Not Bankrupt"
Model predicts the company is not at risk of bankruptcy.

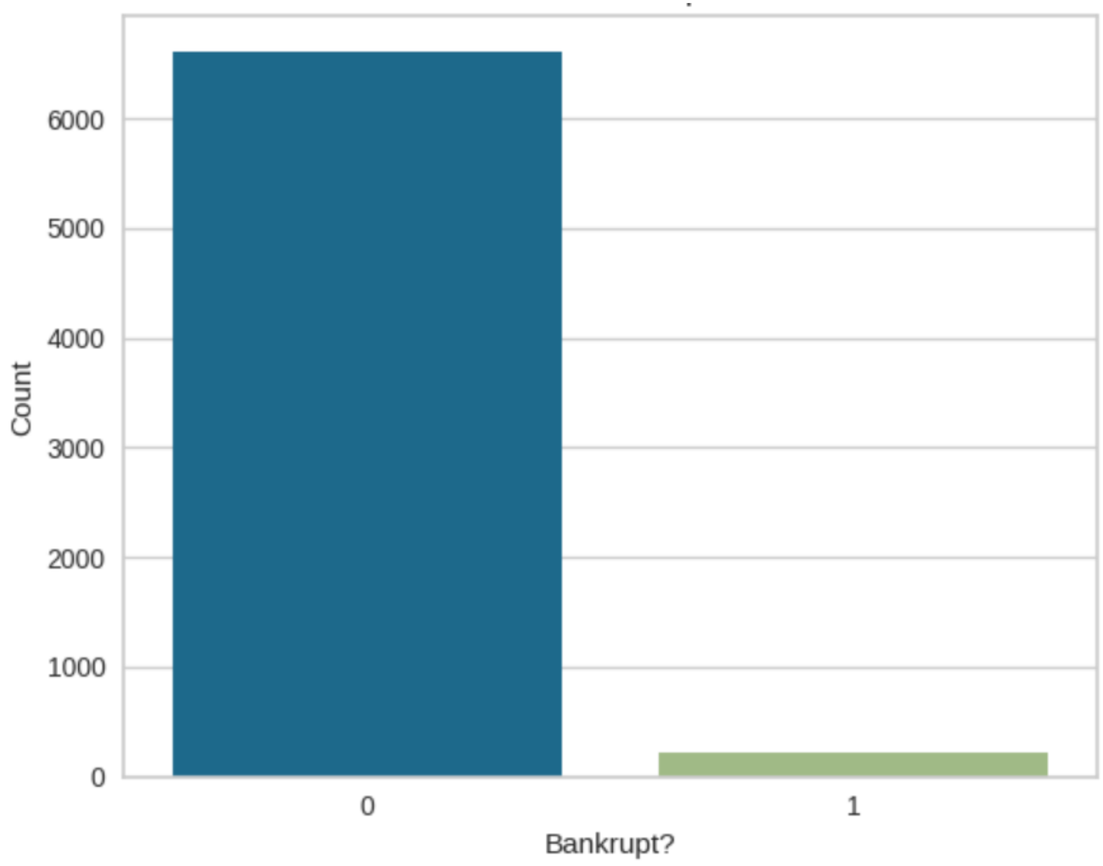


Data preprocessing

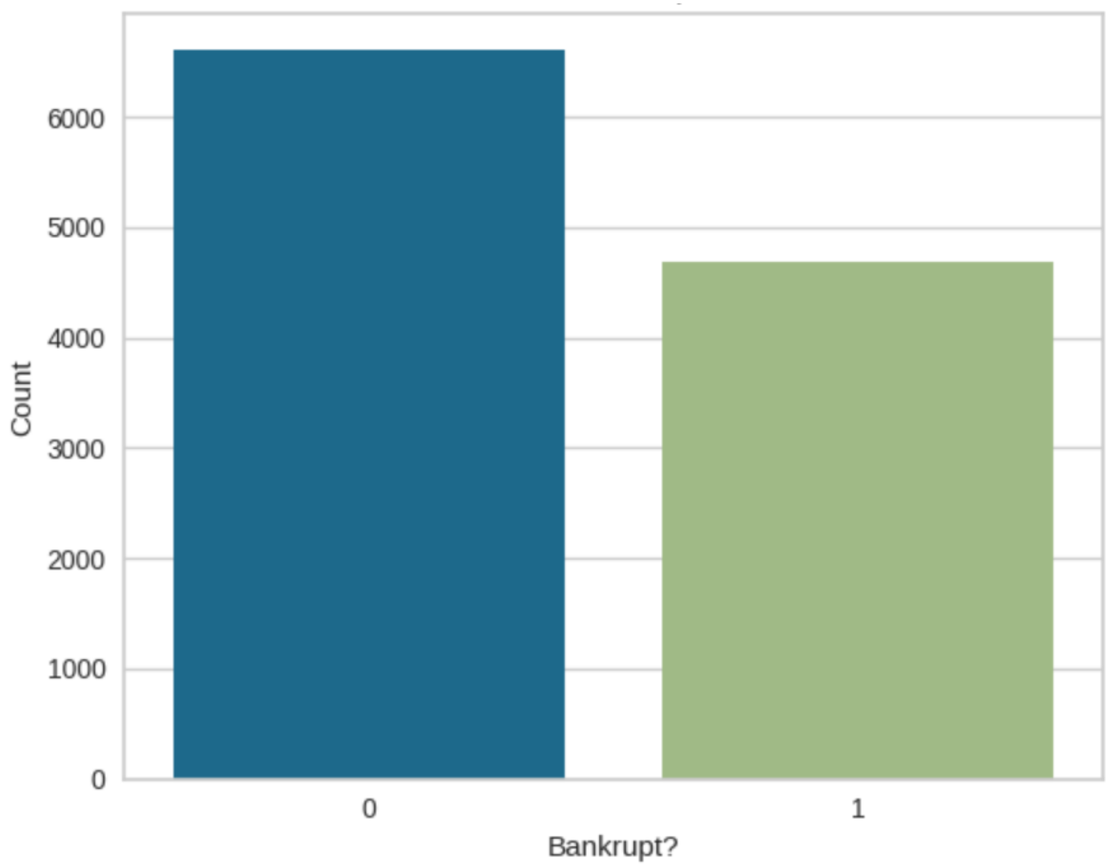
Imbalanced dataset



Addressing Imbalance in the Dataset



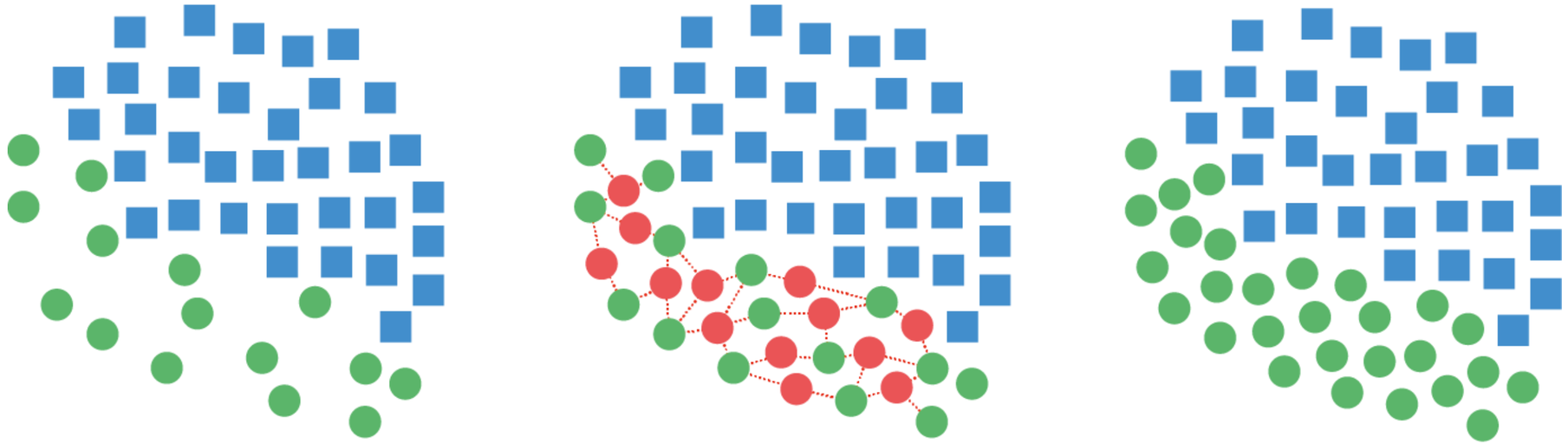
Before



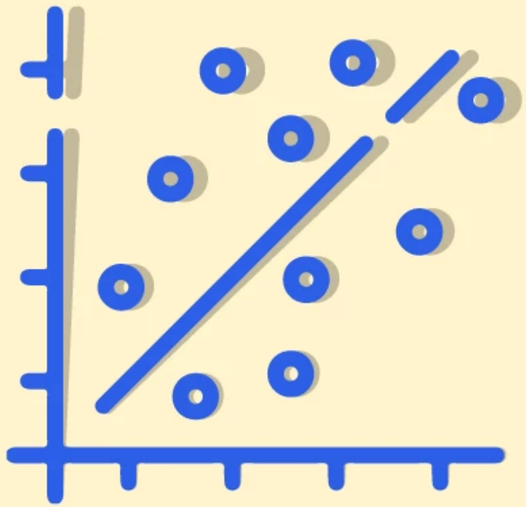
After

SMOTE technique

Synthetic Minority Oversampling Technique



Correlation



What is correlation?

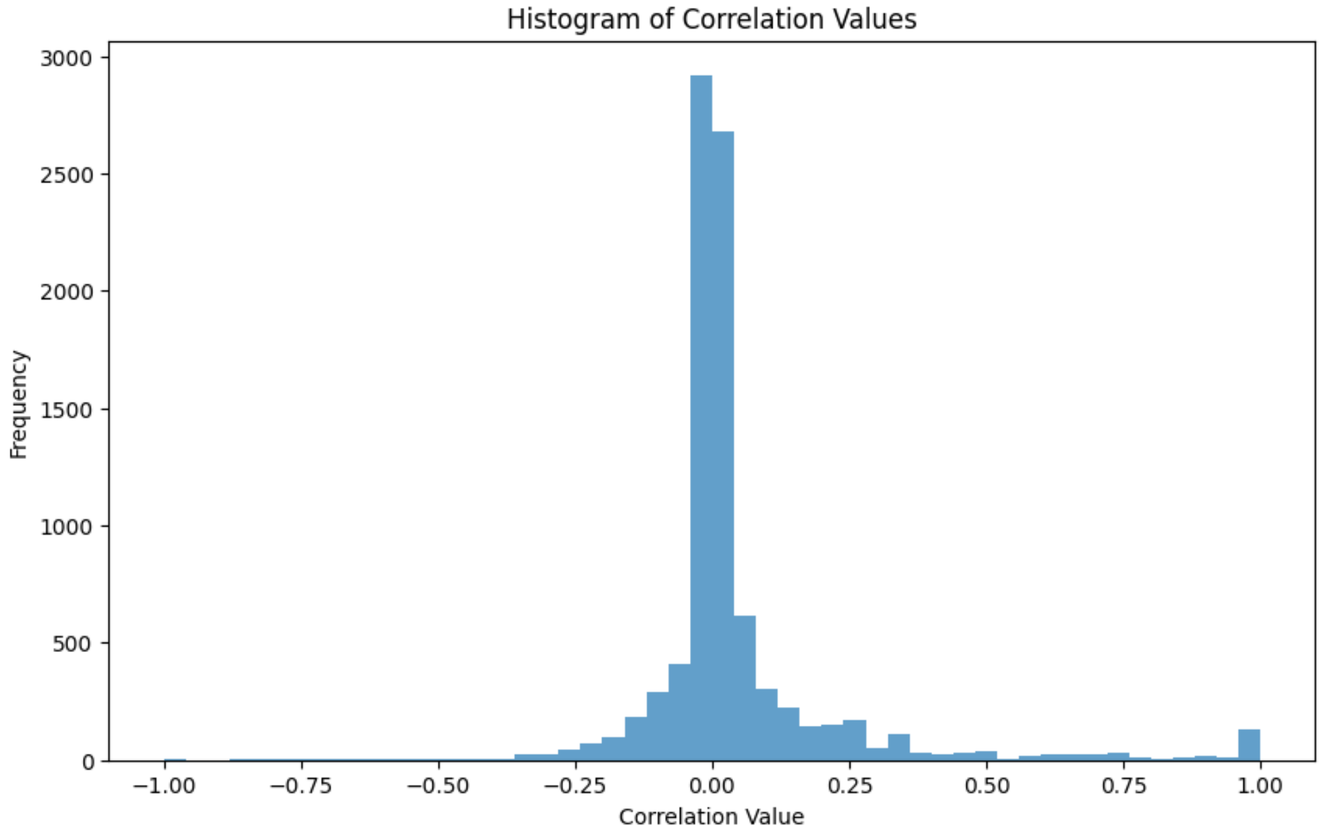
statistical concept that measures the degree to which two variables change together.

Why do we do that?

- provides valuable insights for decision-making
- If two variables are strongly correlated, the value of one variable may provide information about the likely value of the other.

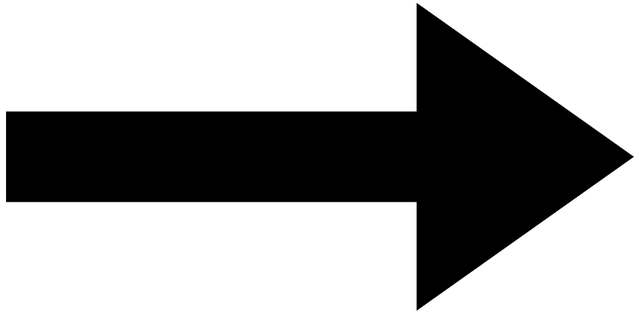
Correlation

treshold:0.75



Multicollinearity Removal

95



64

No missing values found in the dataset!

Models

Effectiveness measures

1

Accuracy
how many cases were correctly classified by the model, both positively and negatively

2

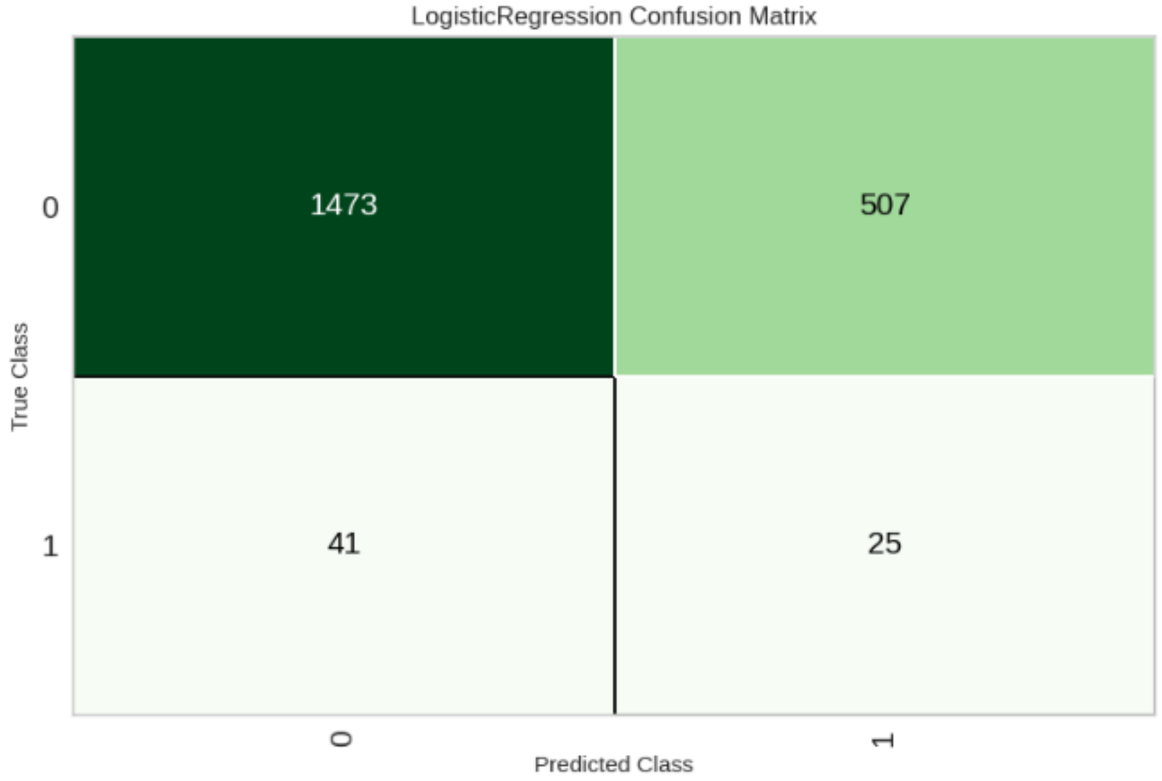
Precision
how many of the predicted positive cases are actually positive

3

Recall
how many actual positive cases were correctly identified by the model

4

F1
implies a good balance between precision and recall.

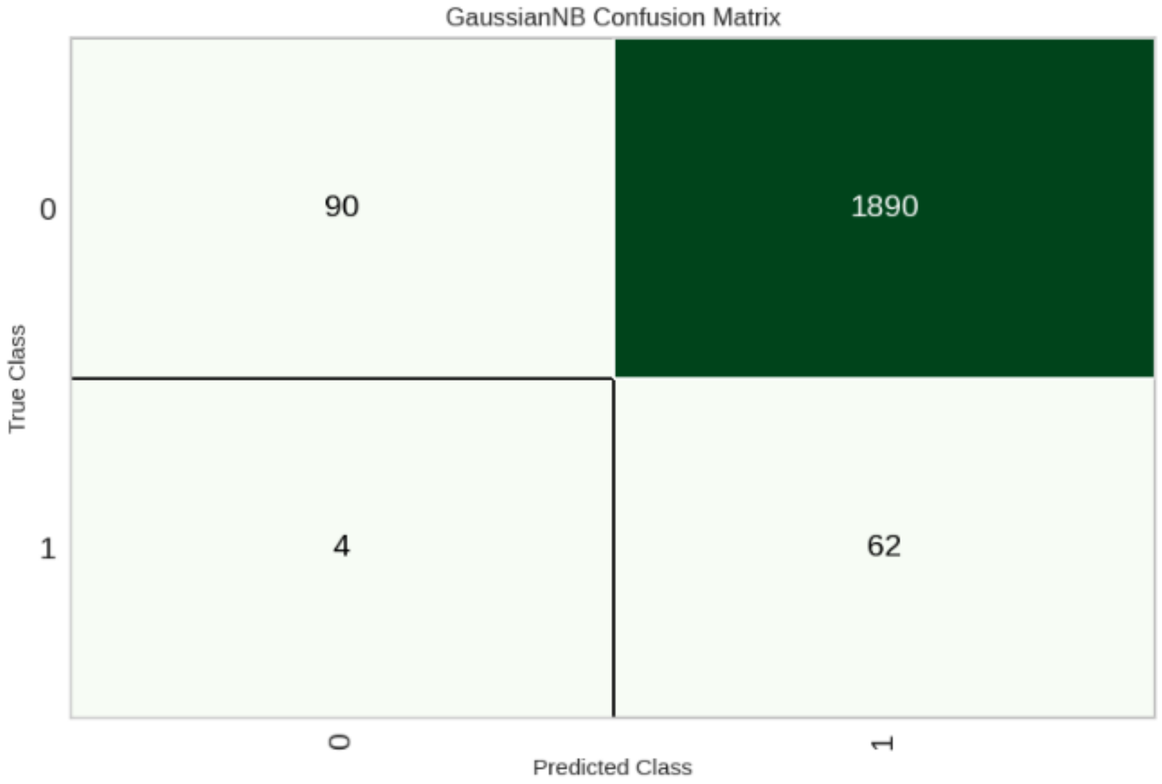


Logistic Regression

Results

Logistic Regression

Accuracy	75.8%
Recall	39%
Precision	5.47%
F1	9.57%

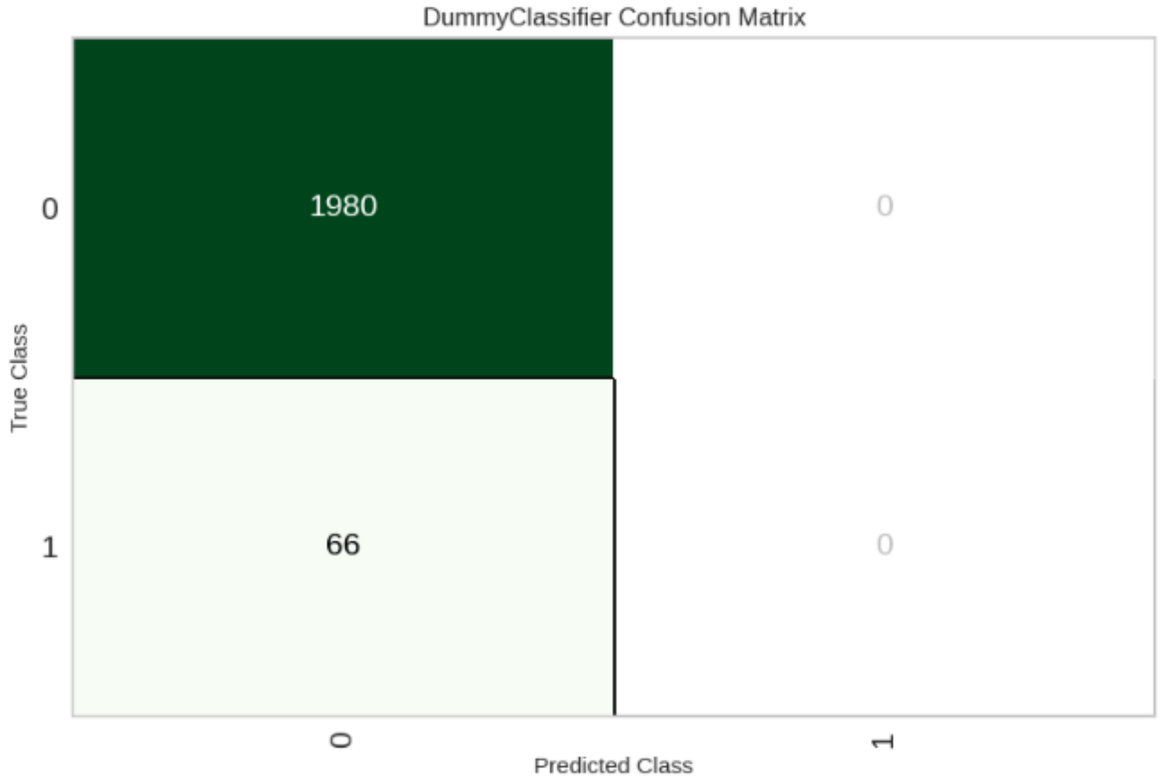


Naive Bayes

Results

Naive Bayes

Accuracy	8.05%
Recall	93.4%
Precision	3.18%
F1	6.15%

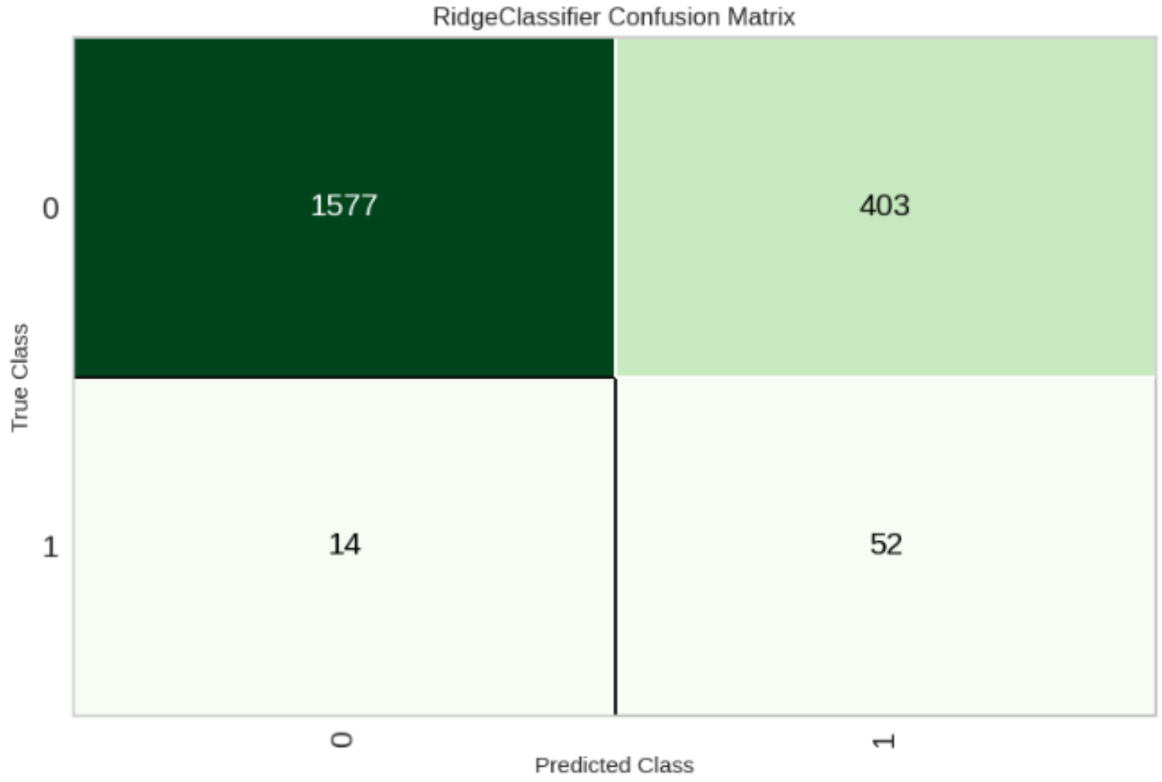


Dummy Classifier

Results

Dummy Classifier

Accuracy	96.7%
Recall	0%
Precision	0%
F1	0%

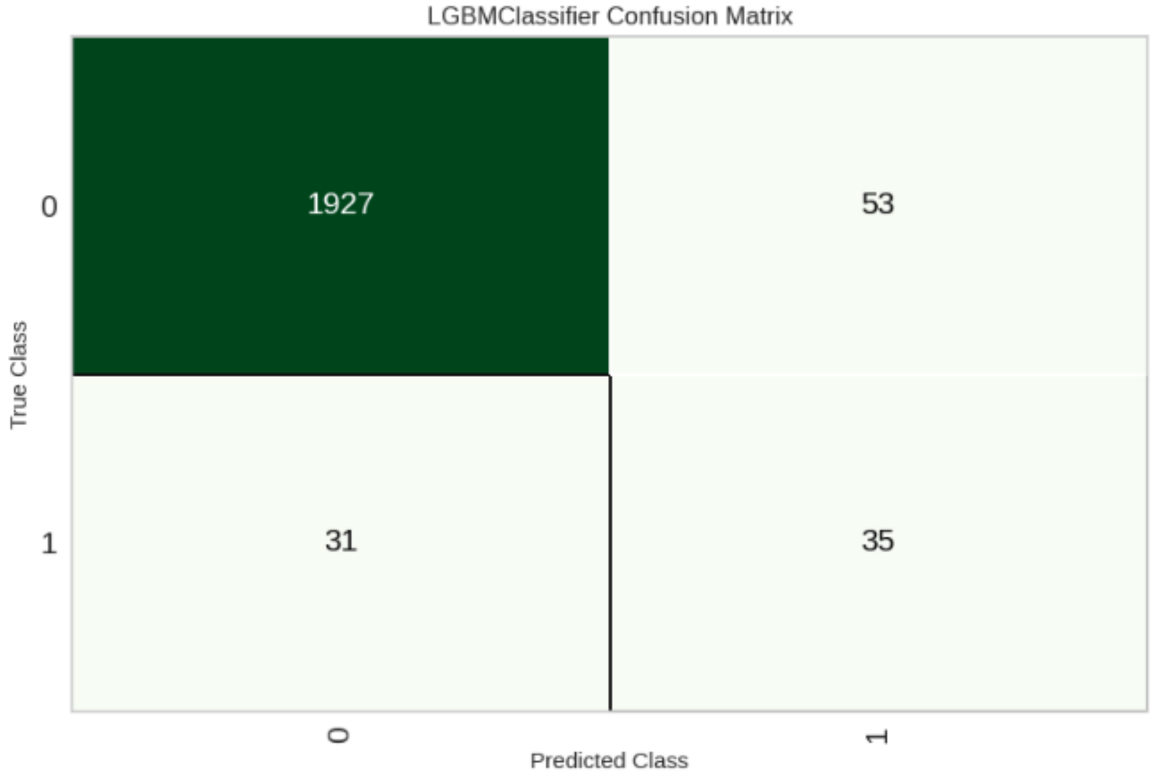


Ridge Classifier

Results

Ridge Classifier

Accuracy	82.5%
Recall	75.9%
Precision	3.18%
F1	6.15%

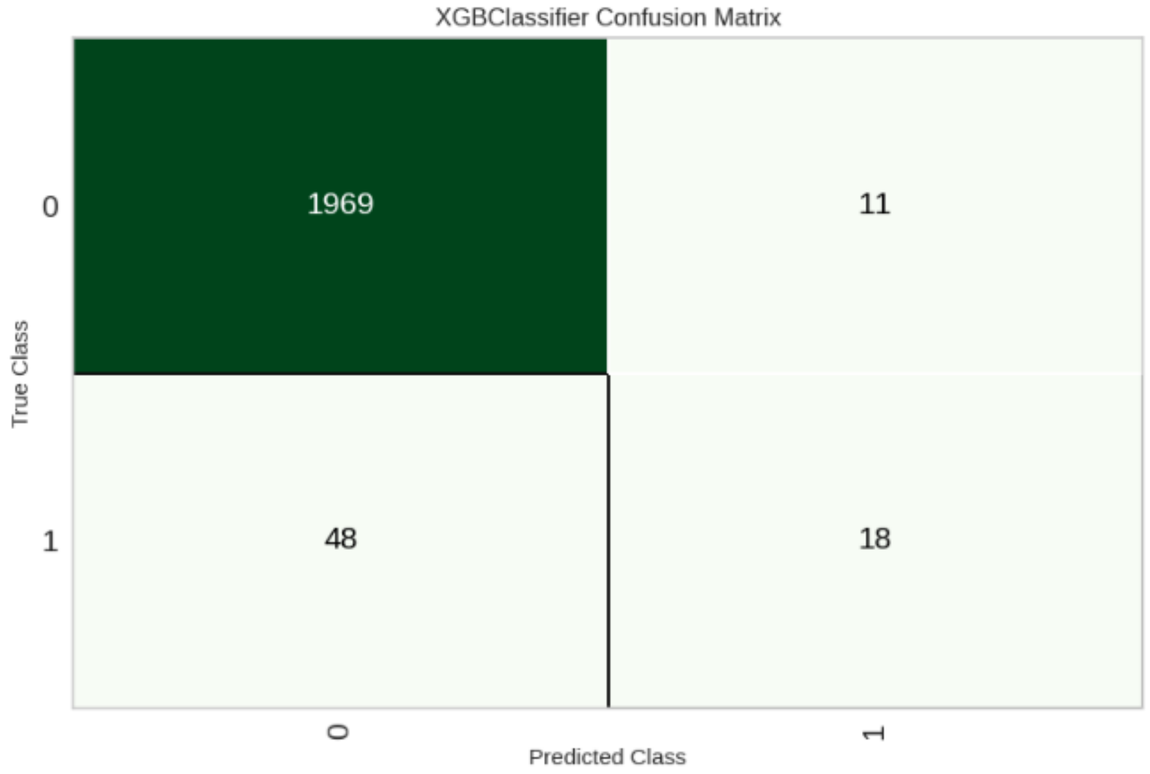


Light Gradient Boosting Machine

Results

Light Gradient Boosting Machine

Accuracy	95.9%
Recall	45.3%
Precision	38.8%
F1	41.5%



Extreme Gradient Boosting

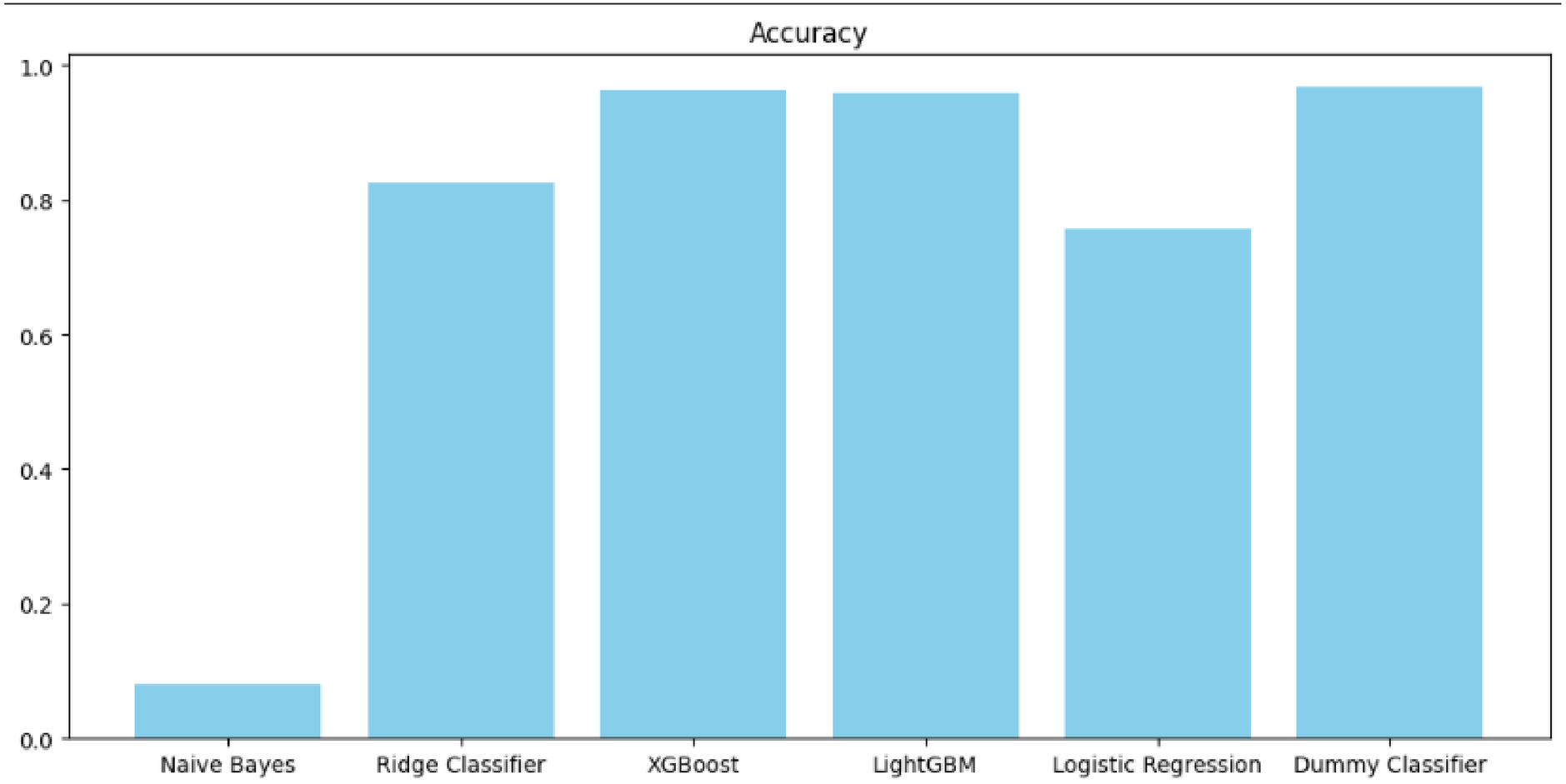
Results

Extreme Gradient Boosting

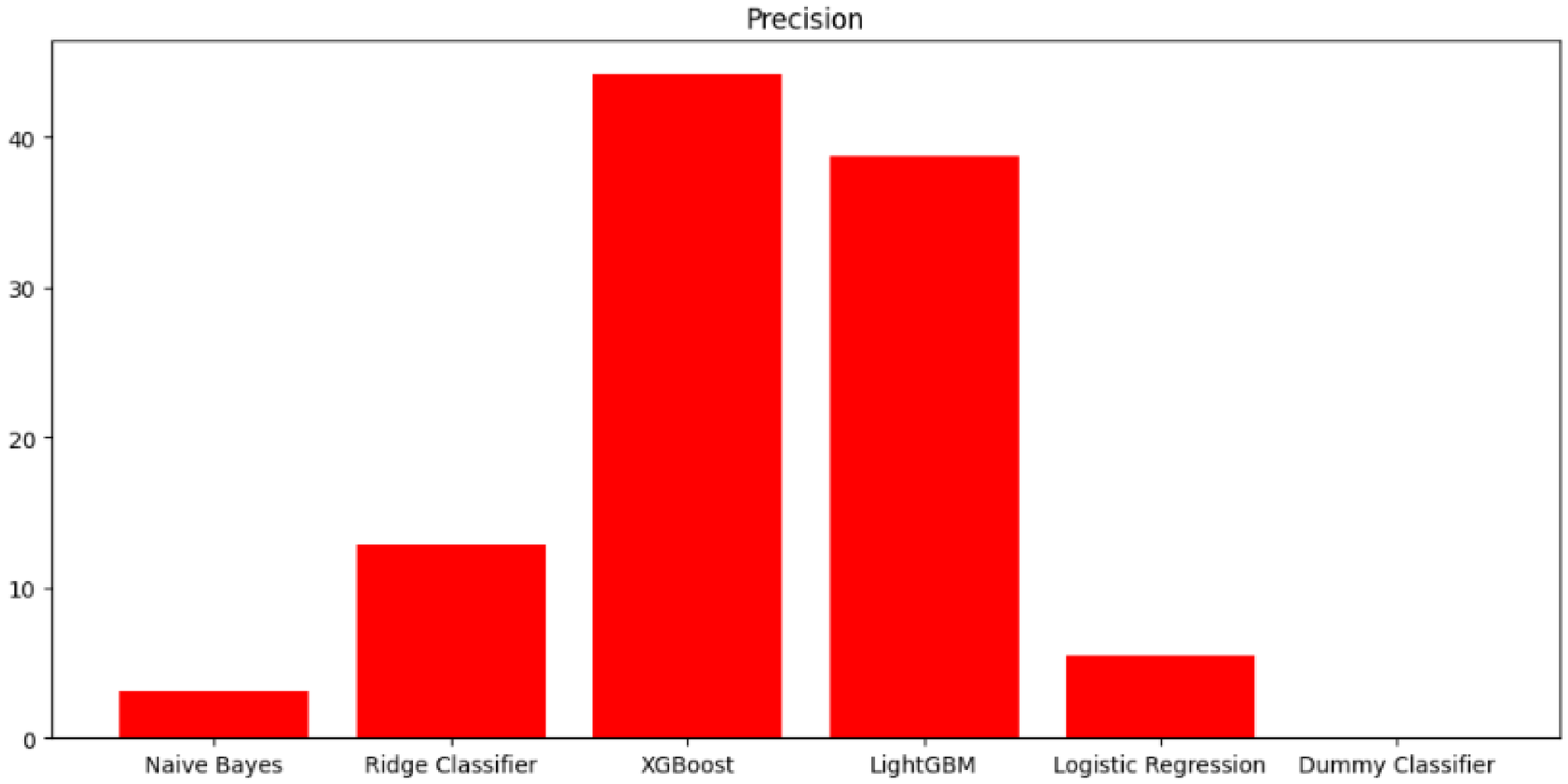
Accuracy	96.3%
Recall	50.5%
Precision	44.2%
F1	46.9%

Testing models' performence

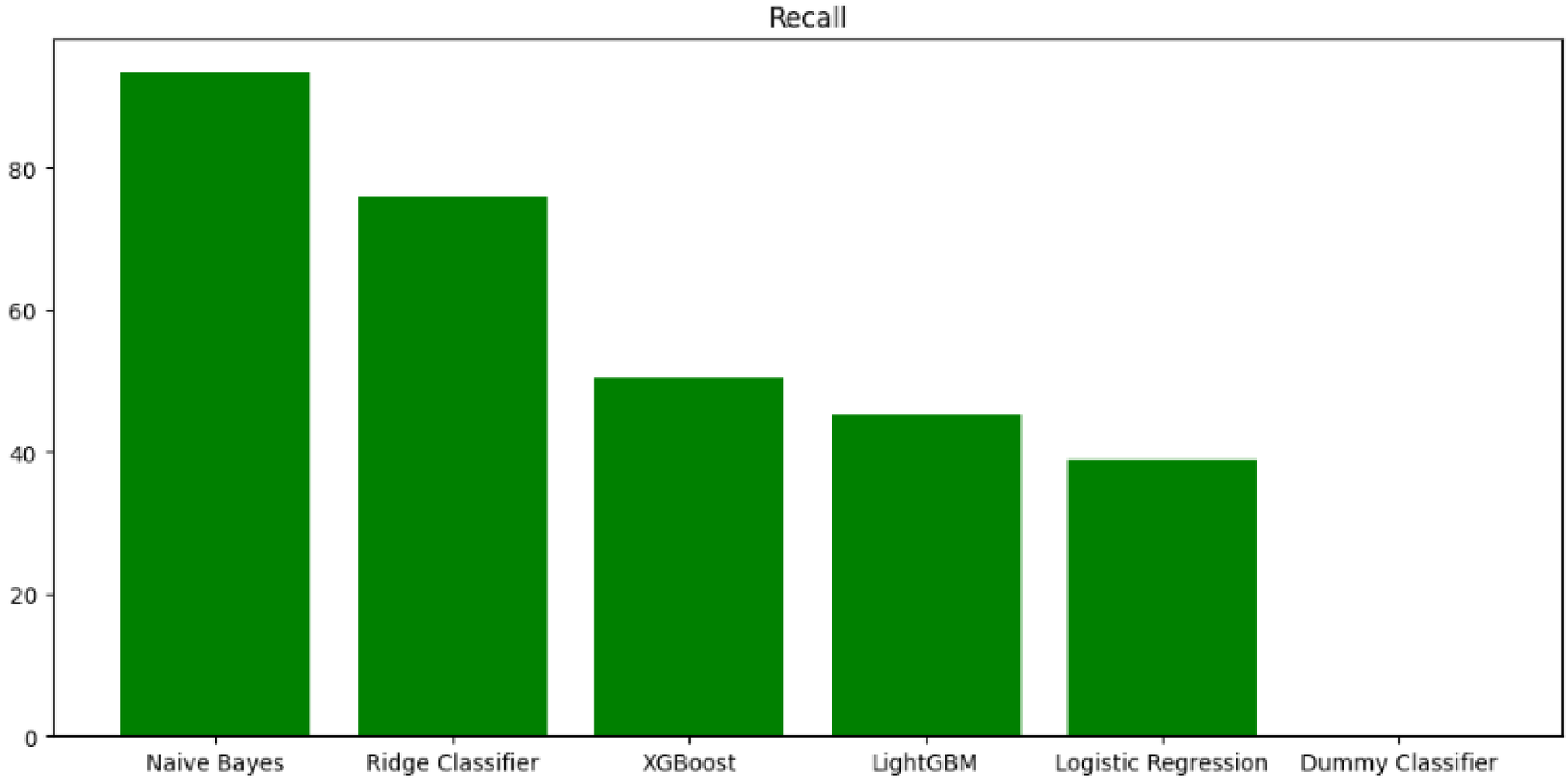
Accuracy Comparasion



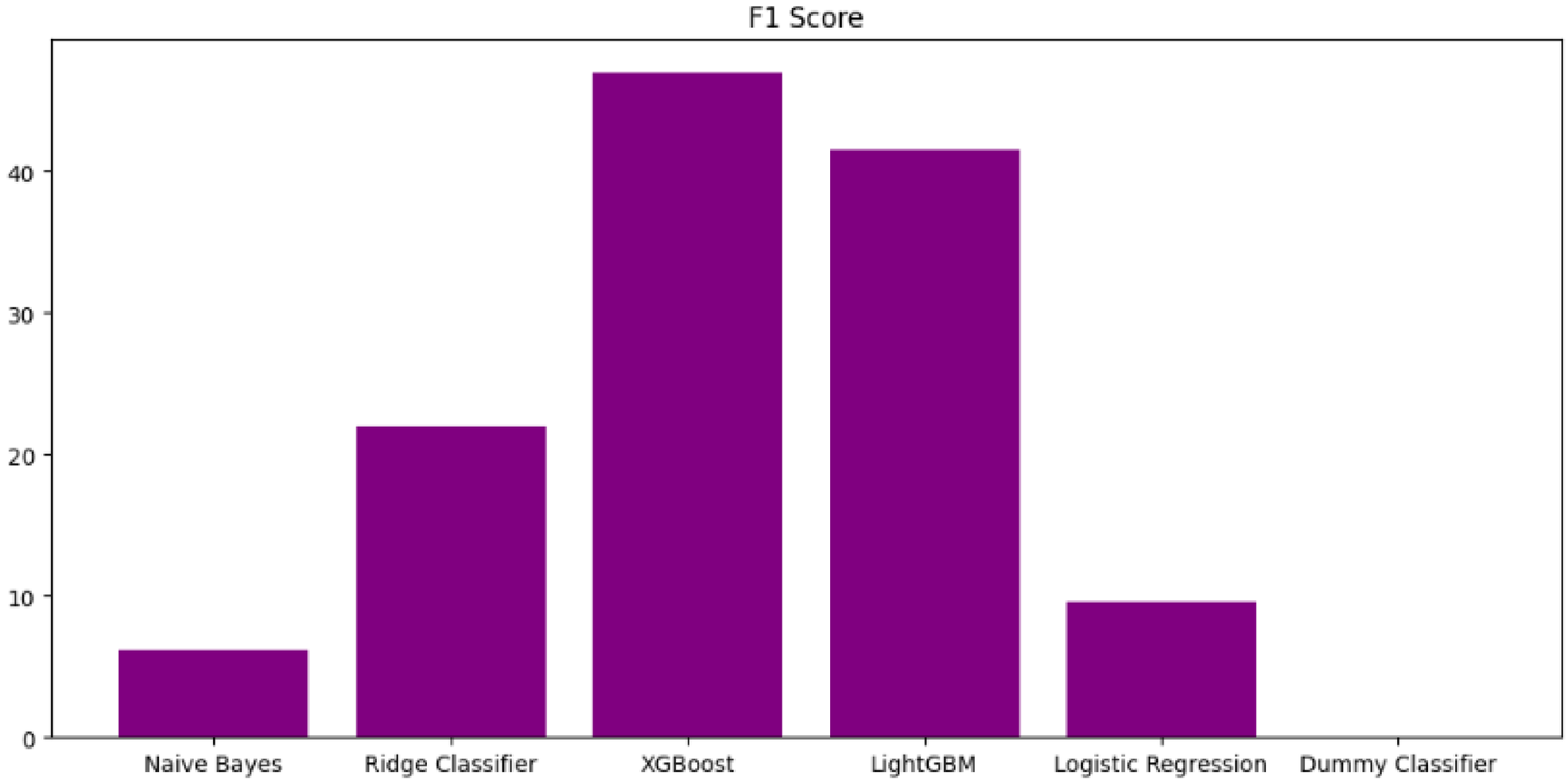
Precision Comparasion



Recall Comparasion



F1 Comparasion



...

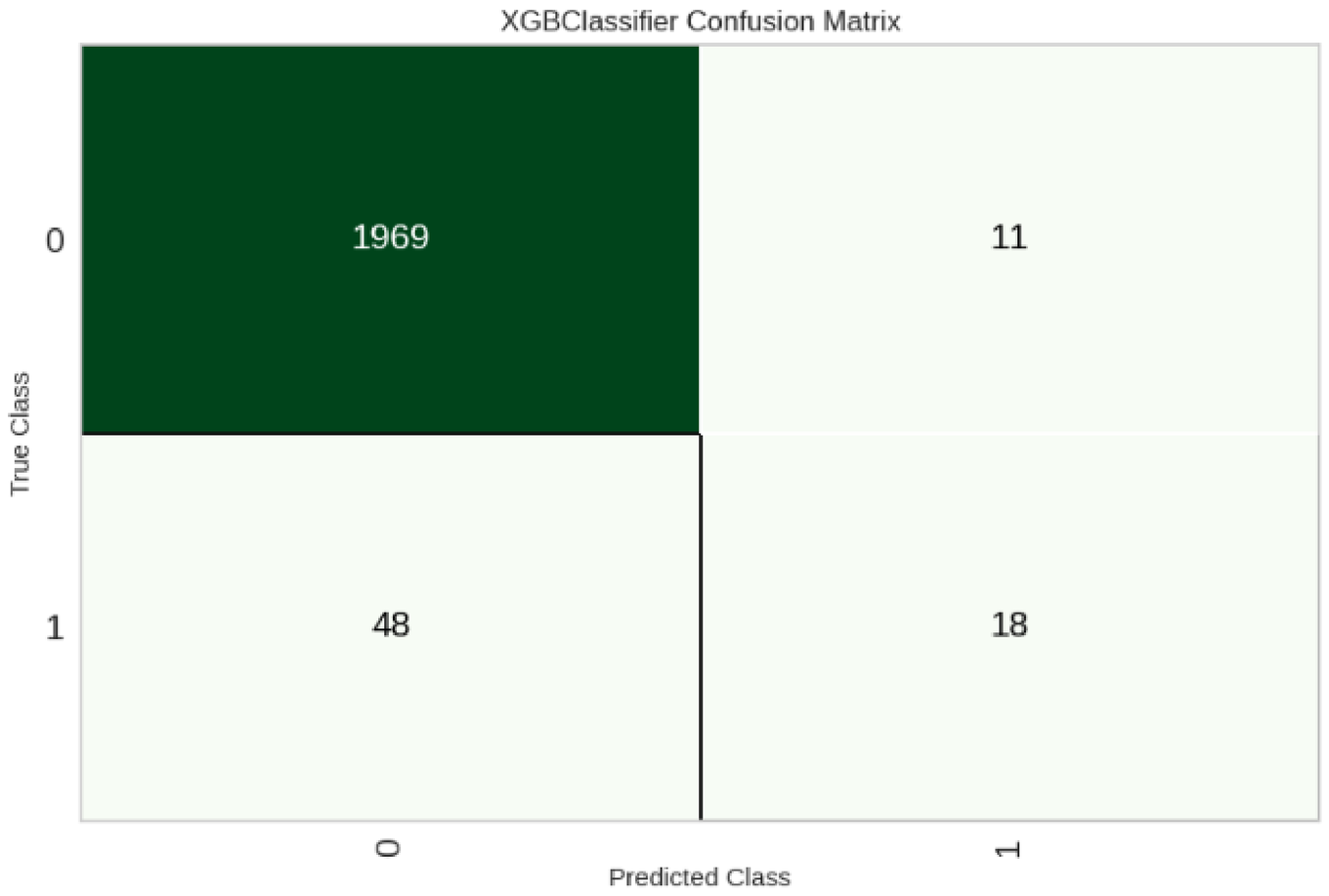
WINNER!

XGBoost

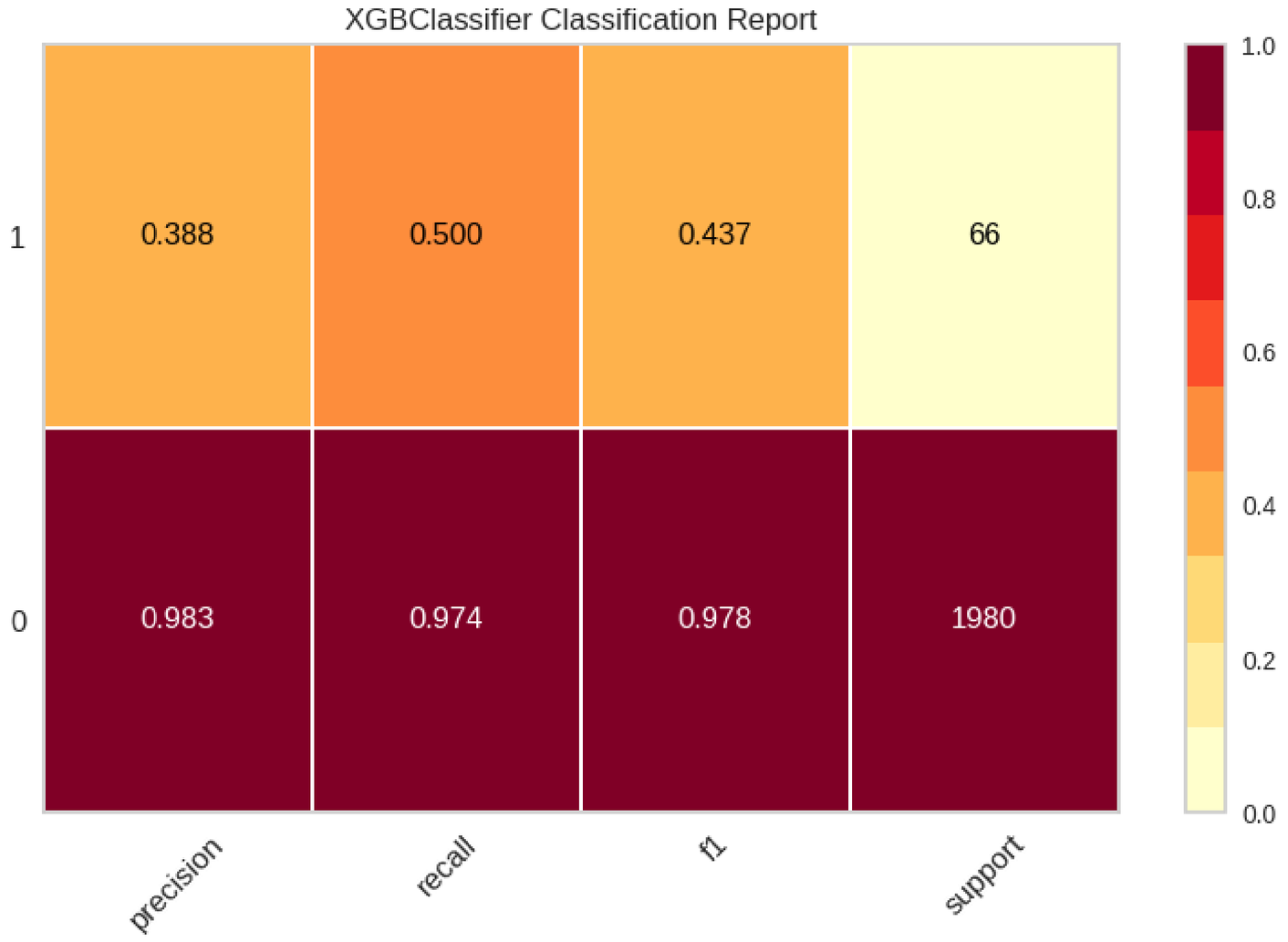
Final results

Accuracy	96.3%
Recall	50.5%
Precision	44.2%
F1	46.9%

Final confusion matrix

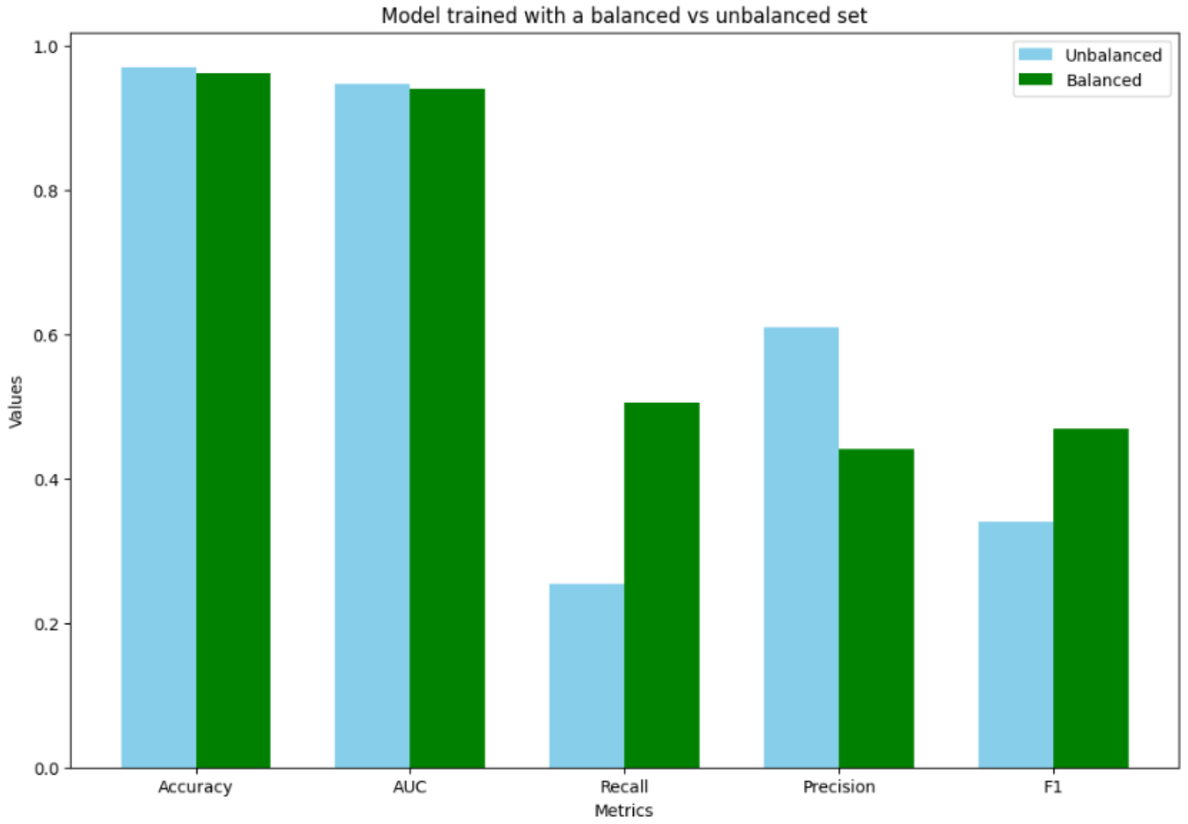


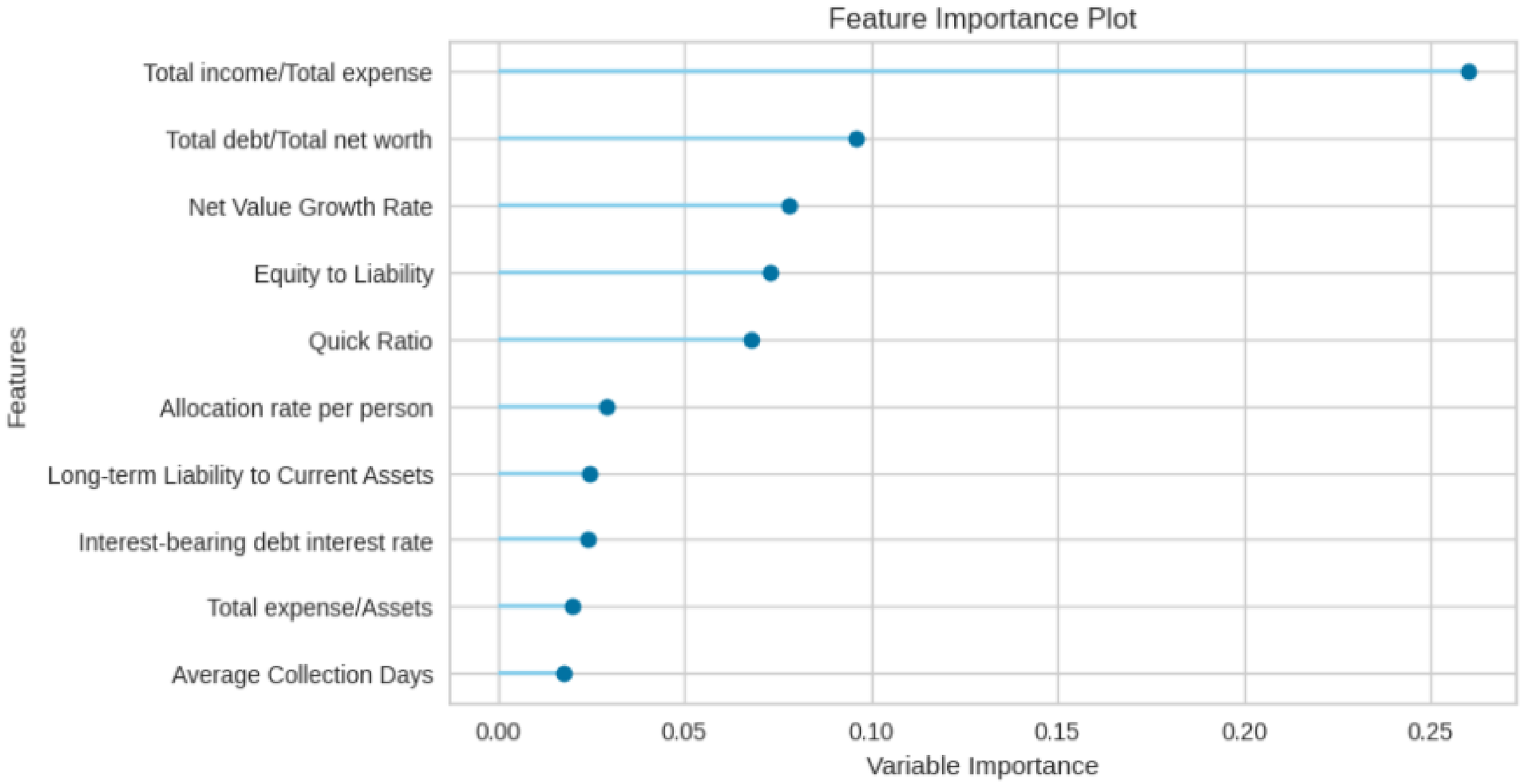
Final classification report



...

The purpose of having a balanced dataset




Feature Importance Plot
Total income/Total expense
Total debt/Total net worth
Net Value Growth Rate
Equity to Liability
Quick Ratio
Allocation rate per person
Long-term Liability to Current Assets
Interest-bearing debt interest rate
Total expense/Assets
Average Collection Days
Features
0.00
0.05
0.10
0.15
0.20
0.25
Variable Importance

Thank you

Nicola Szwaja
Piotr Droś