Bankruptcy Prediction in Poland

Presented By
Ranesh Nair Anil
Sarang Pratap Chamola

Problem Statement

Predict whether a company in Poland will go bankrupt or not.

Polish Companies Bankruptcy

Donated on 4/10/2016

The dataset is about bankruptcy prediction of Polish companies.The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification

Feature Type	# Instances	# Features
Real	10503	65

Dataset Information

Additional Information

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS, http://www.securities.com), which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluate...

SHOW MORE

Has Missing Values?

Yes

Source

The data set is acquired from UCI Machine Learning Repository

Target variable: Class (Binary 0 = Non Bankrupt, 1 = Bankrupt)

43,398

rows

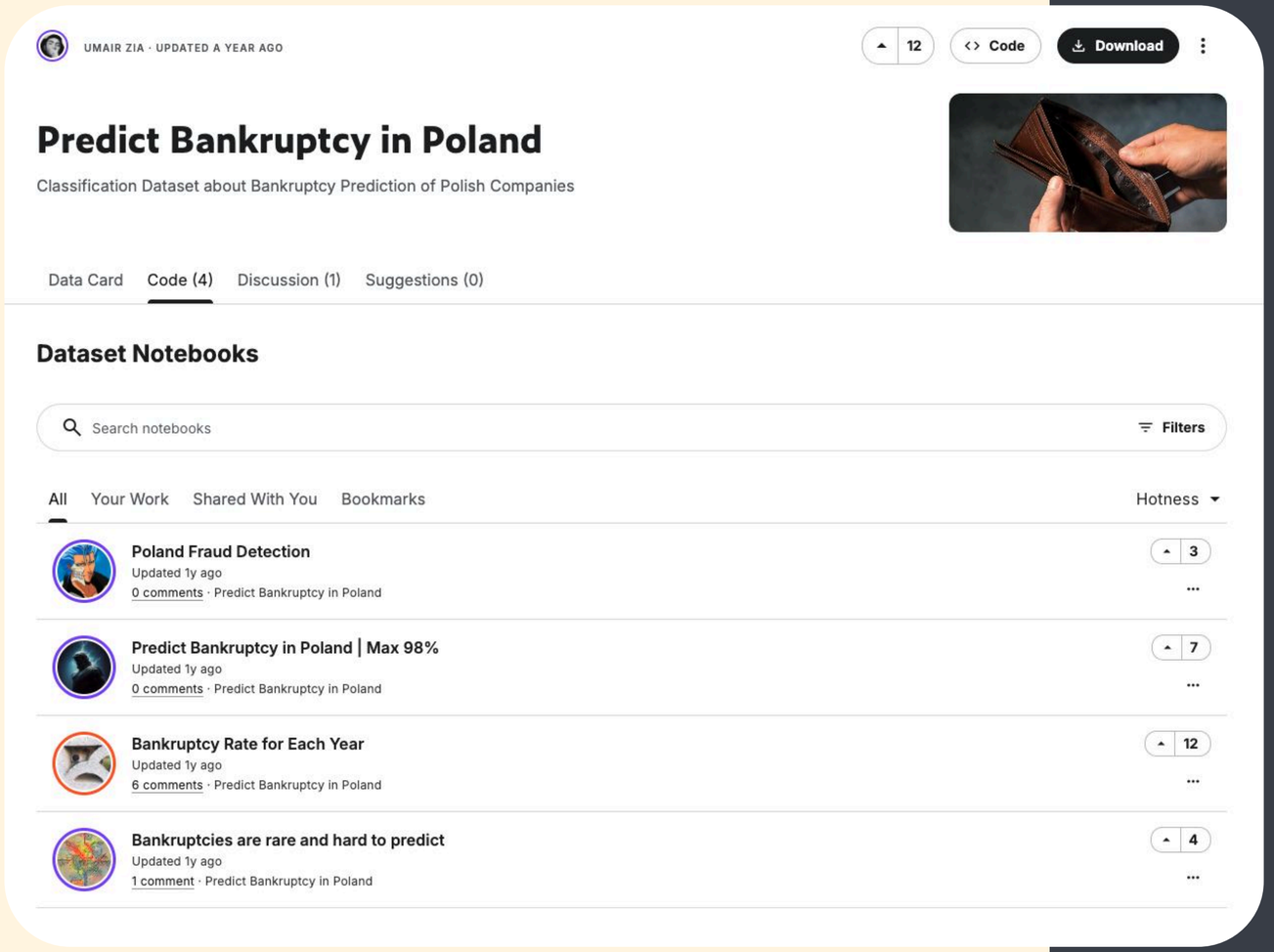
66

columns

The bankrupt companies were analysed in the period 2000-2012 throughout Poland. All features in this dataset are numeric.

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22	A23
0.20055	0.37951	0.39641	2.0472	32.351	0.38825	0.24976	1.3305	1.1389	0.50494	0.24976	0.6598	0.1666	0.24976	497.42	0.73378	2.6349	0.24976	0.14942	43.37	1.2479	0.21402	0.11998
0.20912	0.49988	0.47225	1.9447	14.786	0.0	0.25834	0.99601	1.6996	0.49788	0.26114	0.5168	0.15835	0.25834	677.96	0.53838	2.0005	0.25834	0.152	87.981	1.4293	0.24806	0.12304
0.24866	0.69592	0.26713	1.5548	-1.1523	0.0	0.30906	0.43695	1.309	0.30408	0.31258	0.64184	0.24435	0.30906	794.16	0.45961	1.4369	0.30906	0.2361	73.133	1.4283	0.3026	0.18996
0.081483	0.30734	0.45879	2.4928	51.952	0.14988	0.092704	1.8661	1.0571	0.57353	0.092704	0.30163	0.094257	0.092704	917.01	0.39803	3.2537	0.092704	0.071428	79.788	1.5069	0.1155	0.062782
0.18732	0.61323	0.2296	1.4063	-7.3128	0.18732	0.18732	0.6307	1.1559	0.38677	0.18732	0.33147	0.12182	0.18732	1133.2	0.32211	1.6307	0.18732	0.11553	57.045		0.19832	0.11553
0.22822	0.49794	0.35969	1.7502	-47.717	0.0	0.28139	1.0083	1.9786	0.50206	0.28645	0.58691	0.14812	0.28139	620.14	0.58858	2.0083	0.28139	0.14222	107.26	1.7278	0.28104	0.11535
0.11109	0.64744	0.28971	1.4705	2.5349	0.0	0.11109	0.54454	1.7348	0.35256	0.12575	0.18041	0.30963	0.11109	439.94	0.82965	1.5445	0.11109	0.064036	57.733	0.56811	0.0	0.064036
0.53232	0.027059	0.70554	53.954	299.58	0.0	0.6524	35.957	0.65273	0.97294	0.69394	48.966	1.0602	0.6524	14.272	25.575	36.957	0.6524	0.99949	39.978		0.67129	0.81553
0.00902	0.63202	0.053735	1.1263	-37.842	0.0	0.014434	0.58223	1.3332	0.36798	0.043162	0.033921	0.038938	0.014434	4443.7	0.082138	1.5822	0.014434	0.010827	36.623	1.0752	0.030778	0.006766
0.12408	0.83837	0.14204	1.1694	-91.883	0.0	0.15328	0.19279	2.1156	0.16163	0.18454	0.18284	0.075411	0.15328	1918.1	0.1903	1.1928	0.15328	0.072451	109.97	1.5866	0.15819	0.058651
0.24001	0.44355	0.18835	1.44	-21.165	-0.9319	0.24001	1.2545	4.7447	0.55645	0.24227	0.56072	0.1951	0.24001	174.89	2.087	2.2545	0.24001	0.050585	22.027	1.6856	0.0	0.050585
-0.027117	0.11148	0.11989	2.0754	-31.643	-0.084883	-0.0243	7.6741	0.90732	0.85551	-0.0243	-0.21798	0.10164	-0.0243	754.95	0.48347	8.9702	-0.0243	-0.045825	99.502	1.1625	-0.024649	-0.051136
0.26669	0.34994	0.61147	3.0243	43.087	0.55983	0.33207	1.8577	1.1268	0.65006	0.33207	1.0993	0.12047	0.33207	367.04	0.99444	2.8577	0.33207	0.11496	38.183	1.0581	0.30471	0.092322
0.067731	0.19885	0.081562	2.9576	90.606	0.21265	0.078063	4.029	1.257	0.80115	0.078063	1.8736	0.31036	0.078063	926.03	0.39415	5.029	0.078063	0.30912	44.446	1.1848	0.05373	0.2682
-0.029182	0.21131	0.45264	7.5746	57.844	0.010387	-0.034653	3.7324	1.0241	0.78869	-0.034653	-0.50333	0.004191	-0.034653	23292.0	0.015671	4.7324	-0.034653	-0.043861	105.35	0.99083	-0.050624	-0.036936
-0.033801	1.154	-0.20599	0.8215	-74.451	-0.10413	-0.033801	-0.159	0.97767	-0.18349	-0.033801	-0.029291	-0.012483	-0.033801	-14644.0	-0.024924	0.86657	-0.033801	-0.01467	44.032	1.1049	-0.06741	-0.01467
0.27053	0.29913	0.4687	2.5669	73.395	0.72793	0.33619	2.2315	1.2214	0.6675	0.33619	1.1239	0.20684	0.33619	291.52	1.2521	3.343	0.33619	0.18567	32.984	1.1281	0.33518	0.1494
0.028084	0.24231	0.43224	3.0128	47.935	0.021598	0.039729	3.1037	1.0125	0.75206	0.039729	0.185	0.04419	0.039729	1059.3	0.34458	4.127	0.039729	0.021027	36.232	0.78628	0.054953	0.014863
0.20393	0.56037	0.13495	1.2408	3.158	0.0	0.24291	0.78452	2.2706	0.43963	0.24991	0.43347	0.117	0.24291	769.91	0.47408	1.7845	0.24291	0.10698	14.612	1.7474	0.23173	0.089812
0.20876	0.4965	0.42548	2.019	38.934	0.005436	0.25619	1.0141	2.2827	0.5035	0.26624	0.61357	0.14063	0.25619	564.55	0.64654	2.0141	0.25619	0.11223	28.801	1.2501	0.26383	0.091453
0.11119	0.63174	0.24796	2.0	58.154	0.24347	0.13909	0.58294	0.99034	0.36826	0.16821	0.56092	0.14828	0.13909	1570.3	0.23244	1.5829	0.13909	0.14044	37.267		0.15215	0.11227
-0.30505	1.2523	-0.29222	0.71426	-214.91	-0.30505	-0.30505	-0.25344	0.77037	-0.31738	-0.30505	-0.29829	-0.33838	-0.30505	-1620.6	-0.22523	0.79852	-0.30505	-0.36596	147.44		-0.33399	-0.36596
0.12709	0.5305	0.38069	1.7198	-27.618	0.0	0.15611	0.885	2.1322	0.4695	0.17832	0.29516	0.098824	0.15611	918.94	0.3972	1.885	0.15611	0.073216	90.436		0.12362	0.059602
0.12624	0.66286	0.21916	1.3553	21.588	0.020227	0.15633	0.50862	1.713	0.33714	0.15937	0.25346	0.11511	0.15633	1227.0	0.29747	1.5086	0.15633	0.09126	26.912	1.344	0.15293	0.073699
0.29264	0.29252	0.34415	2.327	108.07	0.0	0.36131	2.4185	1.2512	0.70748	0.37144	1.3932	0.30276	0.36131	281.85	1.295	3.4185	0.36131	0.28876	23.793	1.8963	0.35928	0.23389
0.079997	0.23967	0.32124	2.4739	8.1087	0.20743	0.069715	2.5466	1.0556	0.61034	0.069715	0.31987	0.067313	0.069715	742.04	0.49189	4.1724	0.069715	0.039805	59.273	1.1151	0.084917	0.045676
0.20817	0.6024	0.47935	2.1883	69.659	0.0	0.25911	0.66002	1.6923	0.3976	0.26164	0.6423	0.15635	0.25911	831.04	0.43921	1.66	0.25911	0.15311	43.642	1.2443	0.25951	0.12301
0.26345	0.14318	0.61965	5.3277	198.95	0.55971	0.32142	5.8624	1.3437	0.8394	0.32142	2.2448	0.29742	0.32142	147.34	2.4773	6.9841	0.32142	0.26951	44.903	1.3684	0.31094	0.2209
-0.054623	0.92857	-0.026927	0.94263	-87.724	-0.054623	-0.054623	0.074796	0.92301	0.069453	-0.054623	-0.11637	-0.021237	-0.054623	-21123.0	-0.017279	1.0769	-0.054623	-0.072298	77.569		-0.062889	-0.072298
-0.35884	1.1171	-0.46577	0.46115	-152.21	-1.1752	-0.35884	-0.11206	0.97154	-0.12518	-0.35884	-0.41514	-0.19052	-0.35884	-1339.2	-0.27255	0.89517	-0.35884	-0.22455	40.657	1.0345	-0.1916	-0.22455
0.10248	0.28993	0.07732	1.2981	-2.921	0.22034	0.12731	2.095	1.0805	0.60739	0.12731	0.4909	0.10245	0.12731	657.45	0.55517	3.4491	0.12731	0.081028	17.626	2.3868	0.13335	0.065225
0.1409	0.7556	0.33886	1.6388	-11.094	0.36762	0.1409	0.32344	1.0508	0.2444	0.1409	0.26562	0.058424	0.1409	1839.4	0.19844	1.3234	0.1409	0.054902	58.71	1.3822	0.12079	0.054902
0.009297	0.78609	0.07811	1.1433	-97.822	-0.017464	0.01099	0.27211	0.99192	0.21391	0.12421	0.020155	0.020264	0.01099	14275.0	0.02557	1.2721	0.01099	0.011079	117.33	1.2061	0.017641	0.009373
0.001222	0.53108	0.2013	1.5431	-44.797	0.042436	0.002745	0.87355	0.99091	0.46393	0.002745	0.007405	0.033226	0.002745	3841.6	0.095012	1.883	0.002745	0.001807	89.269	1.0293	-0.007323	0.000805
0.013079	0.75841	-0.060128	0.90444	-58.08	0.069955	0.012298	0.29958	0.99307	0.22721	0.012298	0.019544	0.015526	0.012298	7099.8	0.05141	1.3186	0.012298	0.004897	49.665	1.2765	-0.018598	0.005208
0.099873	0.57415	0.1756	1.3169	-1.2568	0.30824	0.11871	0.68902	1.0754	0.3956	0.11871	0.21423	0.094556	0.11871	1449.1	0.25189	1.7417	0.11871	0.077616	18.053	1.1377	0.10364	0.06526
0.053	0.34016	0.64812	3.8148	145.57	0.0	0.35799	1.9398	1.6355	0.65984	0.38165	1.5547	0.24433	0.35799	310.7	1.1748	2.9398	0.35799	0.21889	26.446	1.4323	0.37113	

Prior Gap

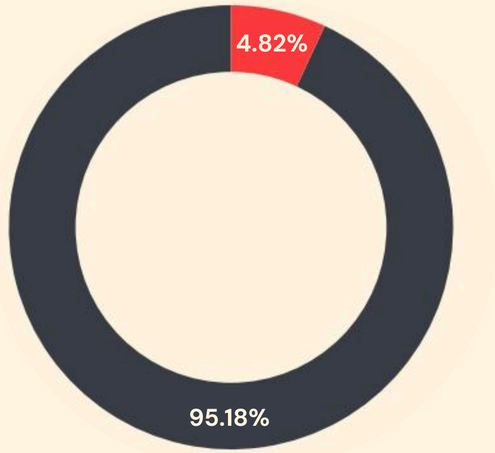


The previous work has compared multiple models and presented their performance scores, which serves as a solid high level analysis. However, while this comparison is useful for identifying the strongest models, it doesn't reveal why certain models perform better.

Our Goal

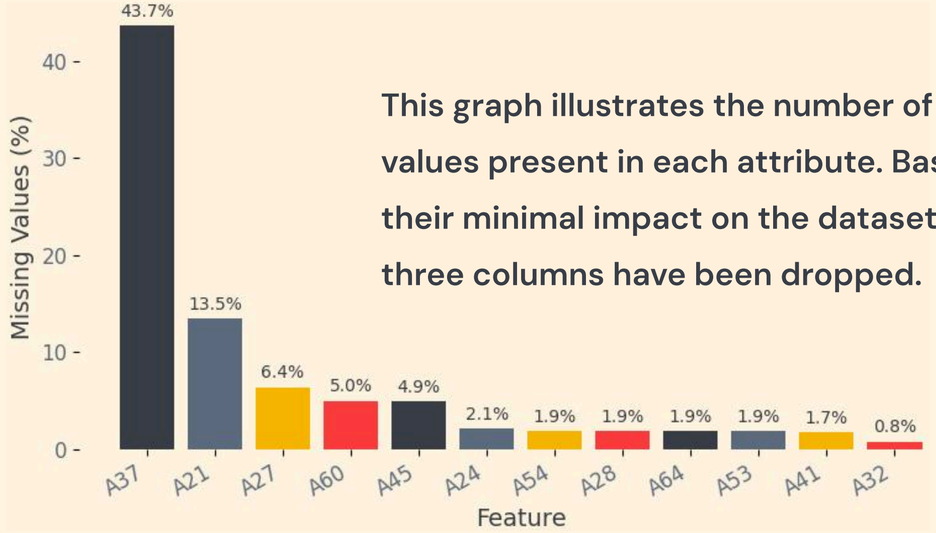
We are conducting a **SHAP analysis**, which goes beyond simple performance metrics by explaining why the model makes certain predictions. SHAP reveals impact of each importance using game theory, making the analysis more transparent, interpretable, and actionable for decision-making.

Imbalanced Data



There are 2,091 instances belonging to class 1 (bankrupt) and 41,314 instances belonging to class 0 (non-bankrupt), indicating a highly imbalanced dataset that will be addressed in later processing.

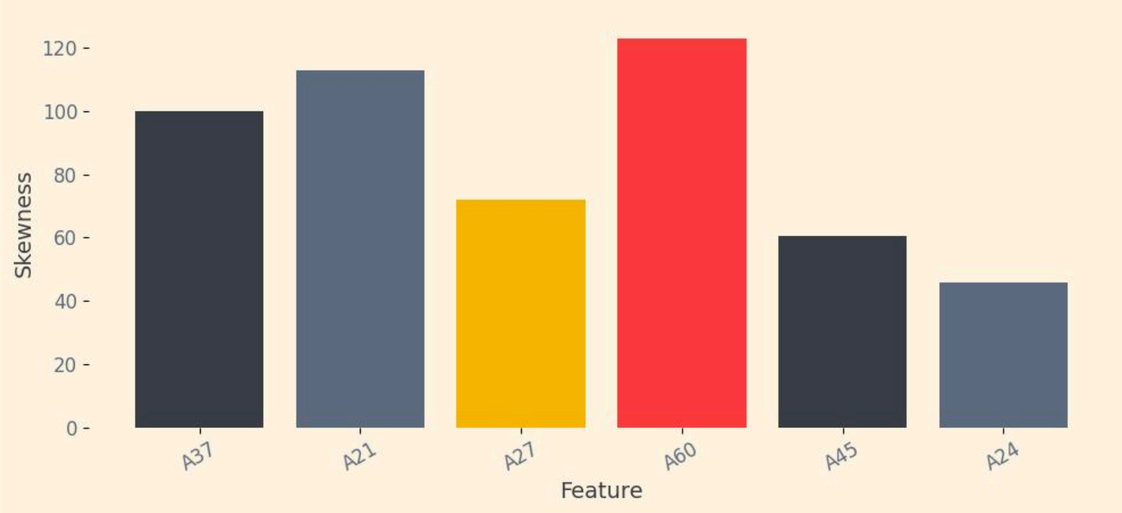
Top 12 missing features by %



This graph illustrates the number of null values present in each attribute. Based on their minimal impact on the dataset, the top three columns have been dropped.

Skewness of the next top missing features

This graph displays the skewness of the next six most important attributes which helped to identify the data distribution.



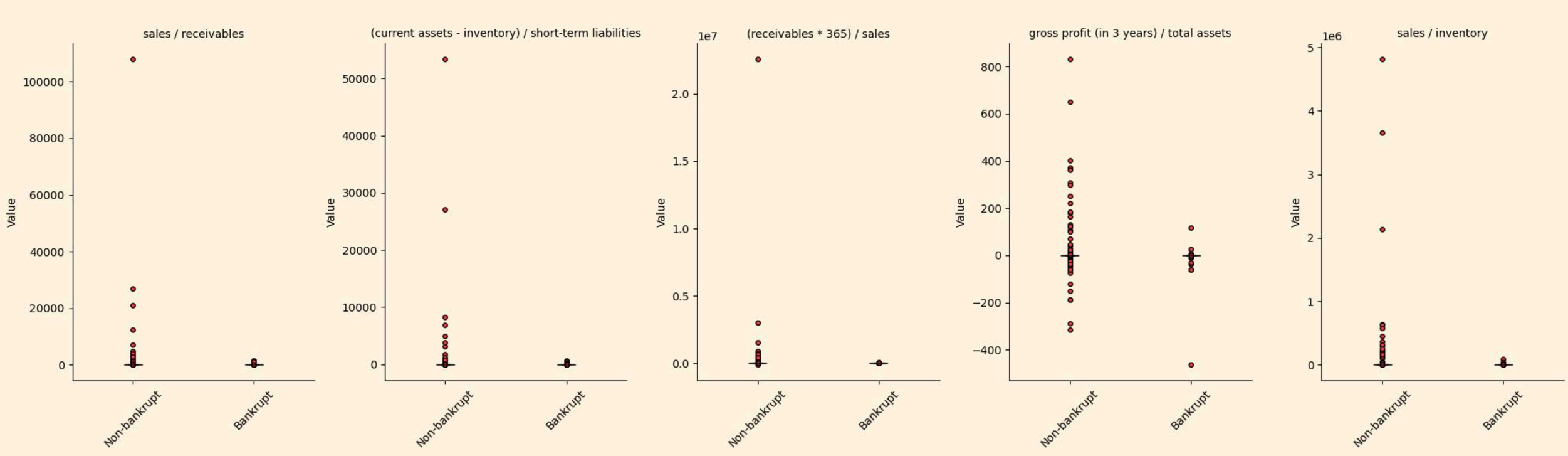
Dropped Features:

A37 - (current assets - inventories) / long-term liabilities

A21 - sales (n) / sales (n-1)

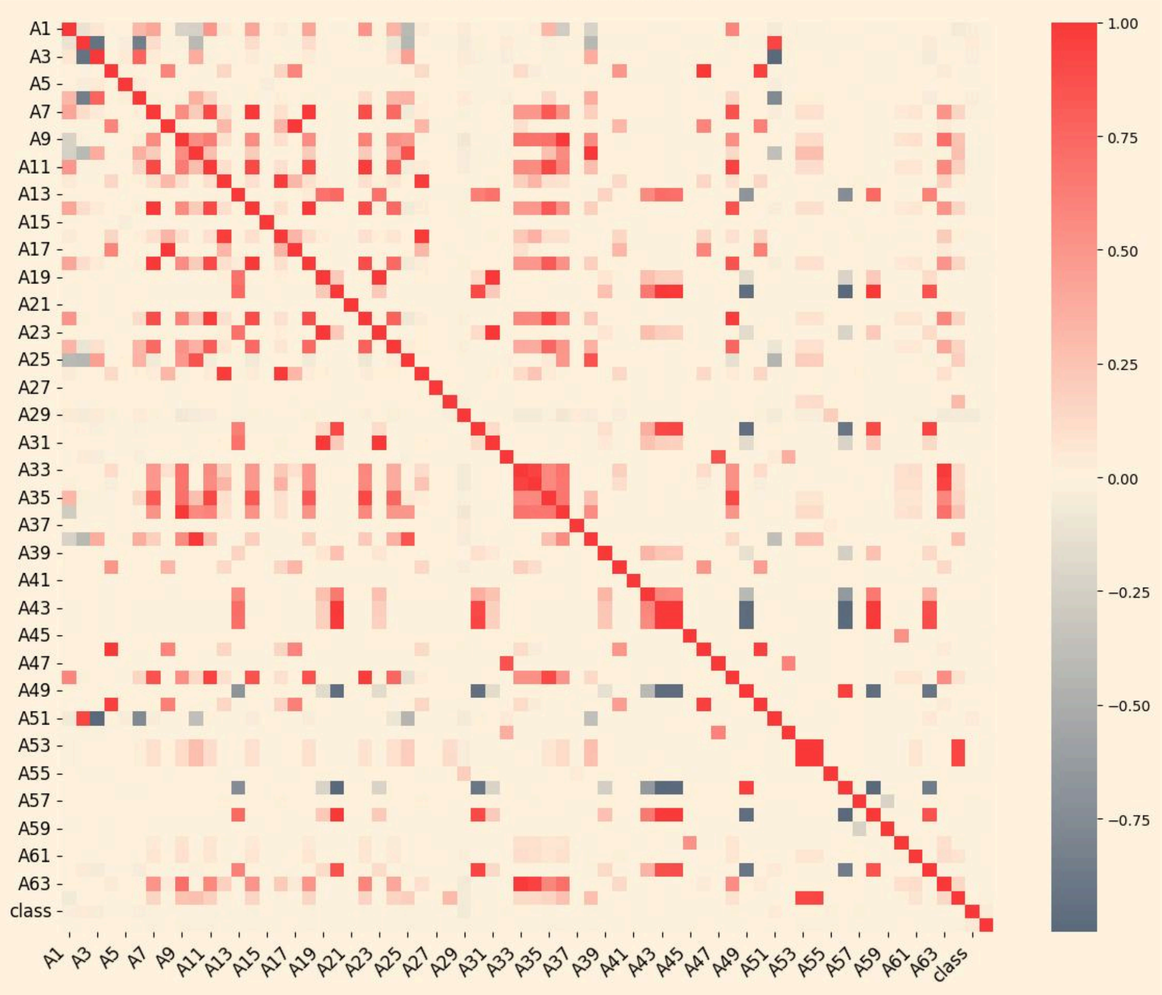
A27 - profit on operating activities / financial expenses

Distribution



Distribution of Top 5 Features by class

Correlation Matrix



Correlation with the Top 20 Features

The correlation heat-map (left) shows the relationships among all features, while the chart of the top 20 features (top) highlights their correlation with the target variable. This helps us understand which features have positive or negative influence on the target.

Attr29 = Logarithm of total assets

Attr3 = Working capital / total assets

Attr2 = total liabilities / total assets

Data Pipeline

CLEANING
filled missing values

SPLIT
85:15 ratio data split

SCALE
removed disproportionalities

SMOTE
on the train data

READY
data is ready for modelling

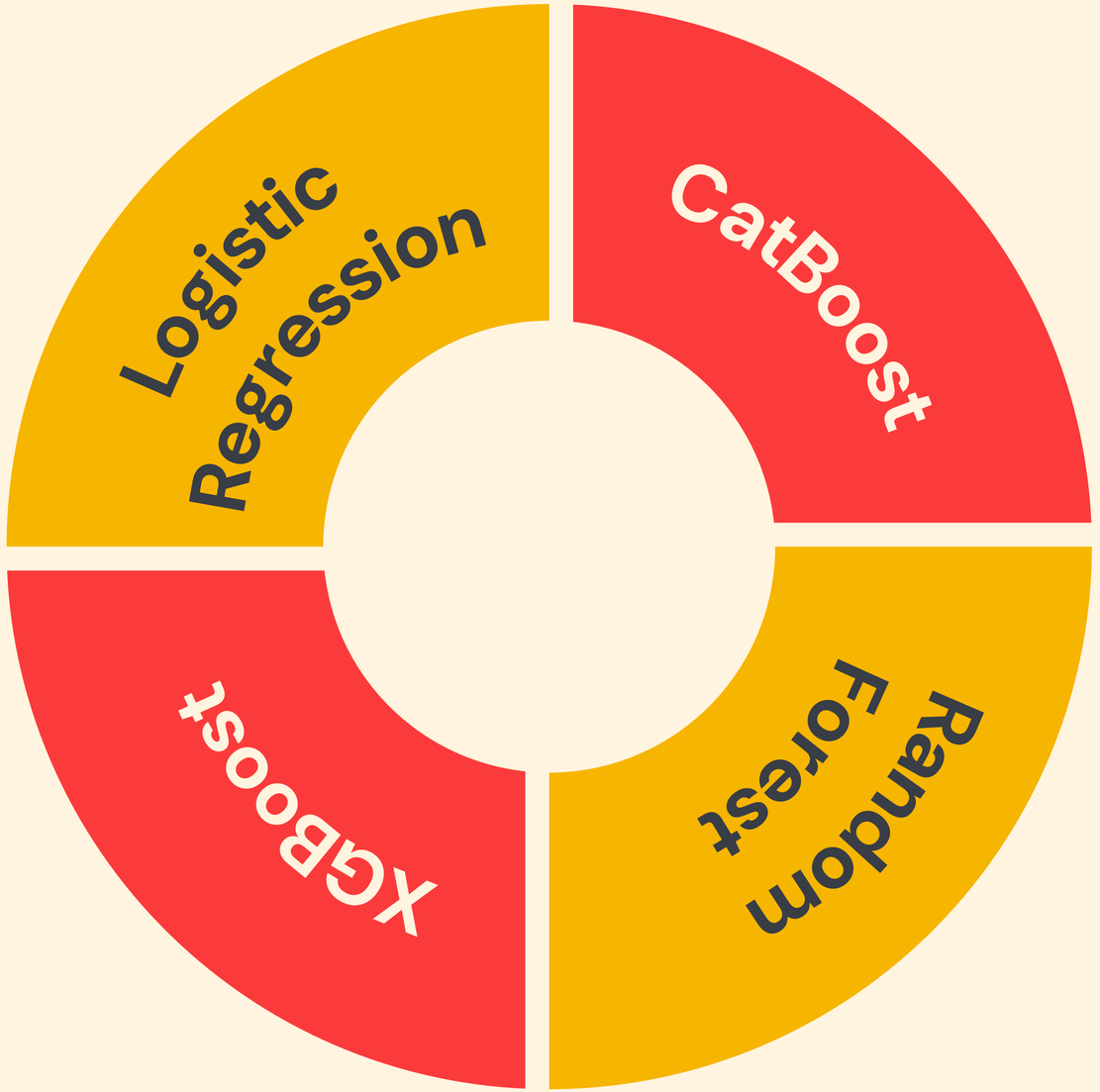
Models Used

Logistic Regression: We have used this linear model to understand the coefficients which show how each financial feature impacts the probability of bankruptcy. It provided us with a strong benchmark to to compare the other models.

Random Forest: This tree enabled model allowed us to understand about the non linear relationships between the financial features and since it is also resistant to outliers and noise it gave us insights.

XGBoost: One of the most powerful ML Algorithms which helped us identify the hard to predict rare bankruptcy cases by using a sequential tree based algorithm.

CatBoost: It naturally handles categorical features without complex encoding and its robust boosting technique reduces overfitting, delivering stable, high-accuracy results with minimal parameter tuning.



Logistic Regression

Classification Report for Logistic Regression Model:

	precision	recall	f1-score	support
Non-bankrupt	0.97	0.72	0.83	6182
Bankrupt	0.11	0.64	0.19	329
accuracy			0.71	6511
macro avg	0.54	0.68	0.51	6511
weighted avg	0.93	0.71	0.79	6511

Cutoff : 0.5

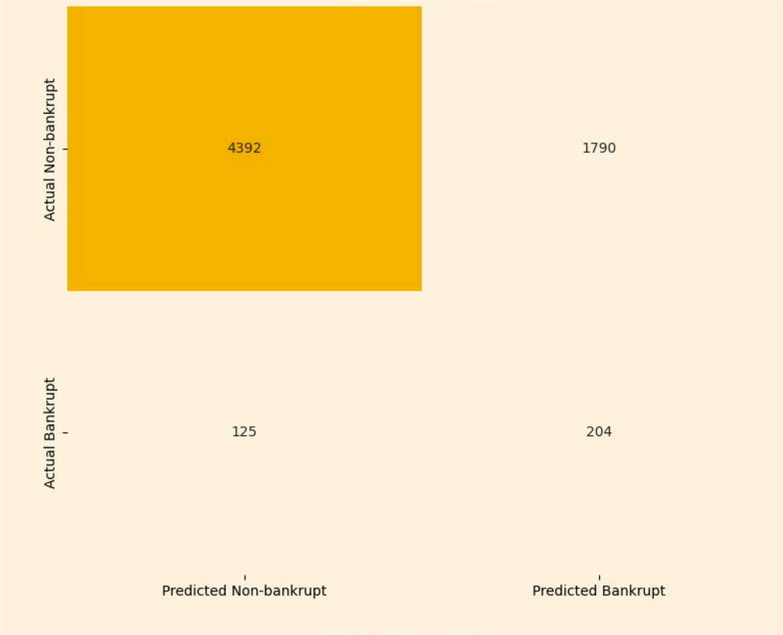
Accuracy = 70.35

AUC = 0.74

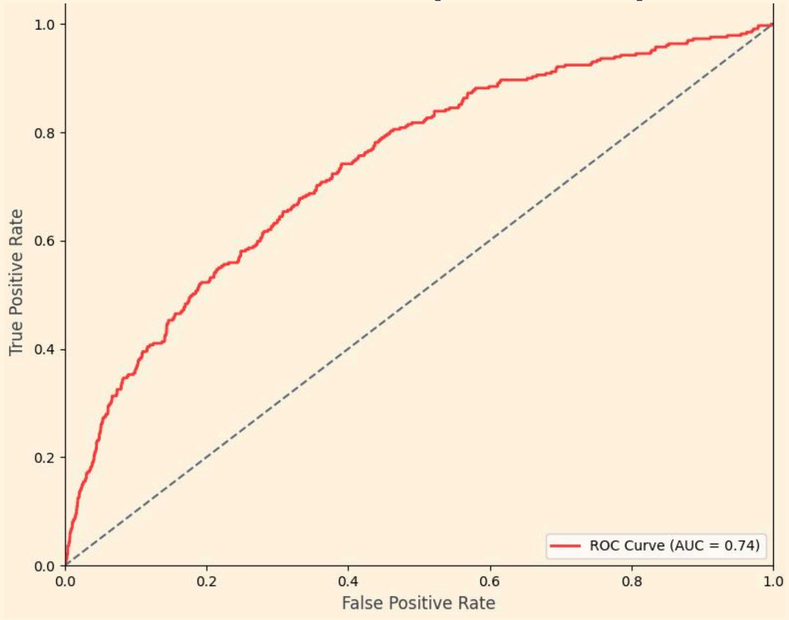
Precision = 0.10

Recall = 0.64

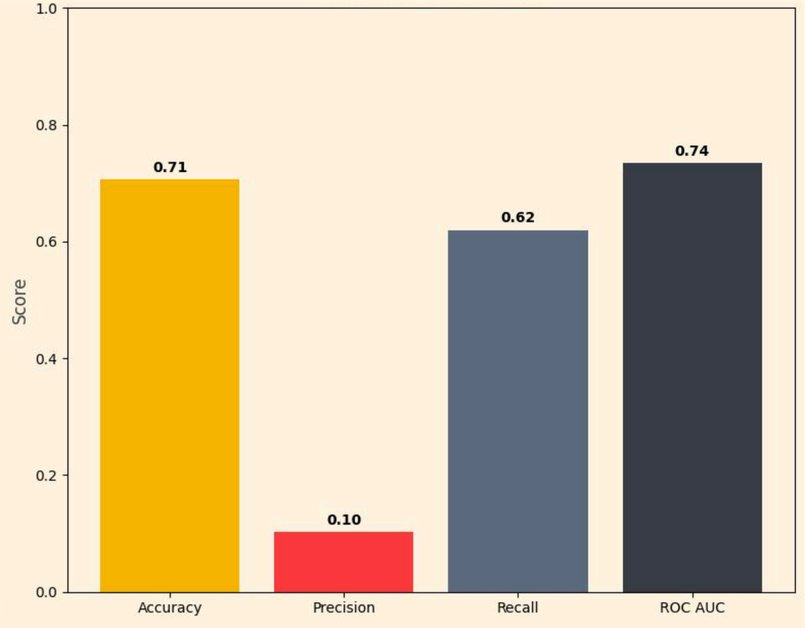
Confusion Matrix



ROC Curve (AUC = 0.74)



Characteristics



Logistic Regression

Bankruptcy ≈

$-0.6107+(15.79\times A_{48})+(8.94\times A_{36})+(3.16\times A_{34})+(3.10\times A_{40})$
$+(2.13\times A_{26})+(2.09\times A_{6})+(1.81\times A_{53})+(1.64\times A_{31})+(1.40\times A_{33})$
$+(1.34\times A_{30})$

Formula:

y_pred = $\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n$

β_0 = intercept

β_i = Coefficients for each feature X_i

Attr	Feature	Co.
A48	EBITDA (profit on operating activities - depreciation) / total assets	15.79
A36	total sales / total assets	8.94
A34	operating expenses / total liabilities	3.16
A40	(current assets - inventory - receivables) / short-term liabilities	3.10
A26	(net profit + depreciation) / total liabilities	2.13
A6	retained earnings / total assets	2.09
A53	equity / fixed assets	1.81
A31	(gross profit + interest) / sales	1.64
A33	operating expenses / short-term liabilities	1.40
A30	(total liabilities - cash) / sales	1.34

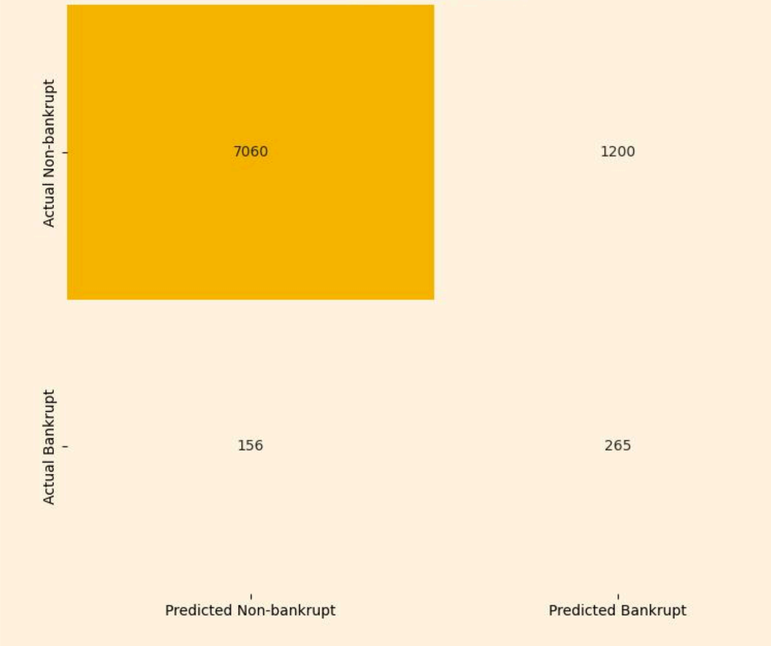
Random Forest

```
Classification Report for Random Forest Model:
              precision    recall  f1-score   support

Non-bankrupt       0.97      0.97      0.97      8260
    Bankrupt       0.34      0.32      0.33       421

    accuracy                           0.94      8681
   macro avg       0.65      0.65      0.65      8681
weighted avg       0.94      0.94      0.94      8681
```

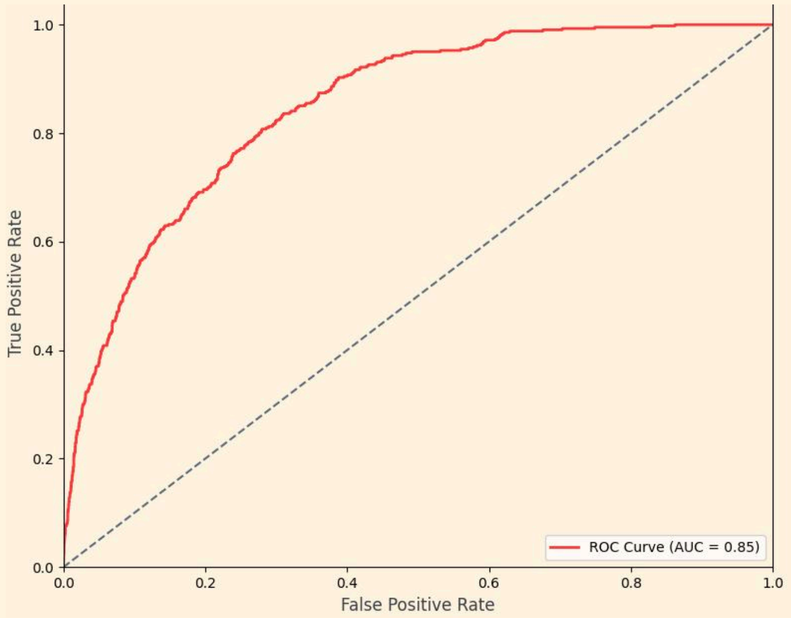
Cutoff : 0.65

- Accuracy = 92.82
- AUC = 0.85
- Precision = 0.31
- Recall = 0.40

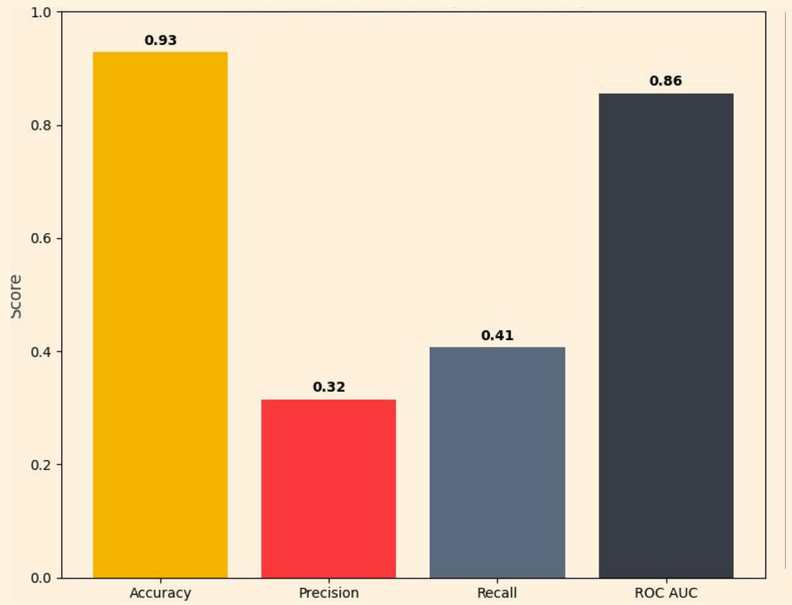
Confusion Matrix



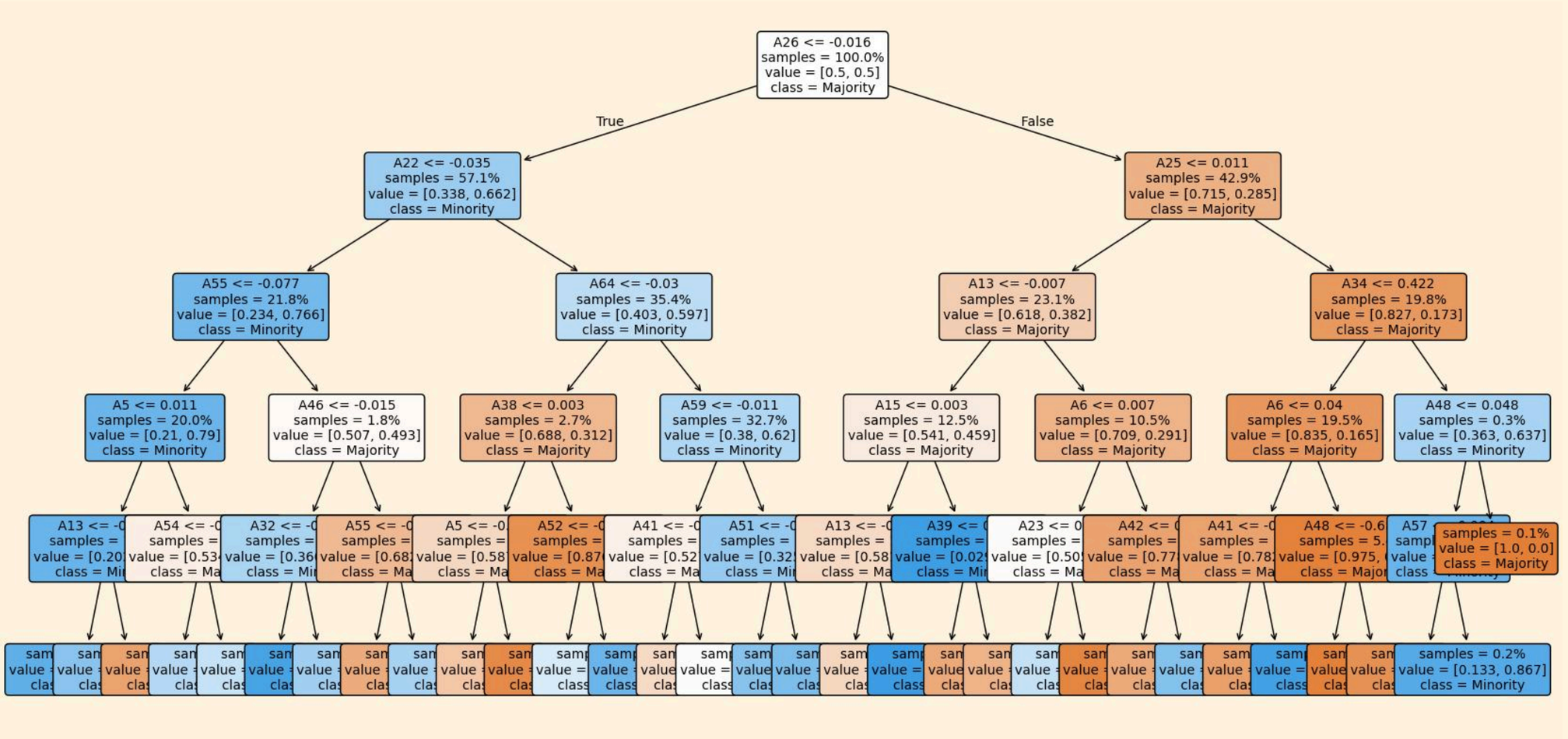
ROC Curve (AUC = 0.74)



Characteristics



Random Forest




```
Fold 1
  Fold 1 Metrics:
    Accuracy:  0.9297
    Precision: 0.3346
    Recall:    0.3878
    AUC:       0.8538
    Best Threshold used: 0.675
Fold 2
  Fold 2 Metrics:
    Accuracy:  0.9196
    Precision: 0.2897
    Recall:    0.4699
    AUC:       0.8614
    Best Threshold used: 0.630
Fold 3
  Fold 3 Metrics:
    Accuracy:  0.9349
    Precision: 0.2989
    Recall:    0.3677
    AUC:       0.8504
    Best Threshold used: 0.681
Fold 4
  Fold 4 Metrics:
    Accuracy:  0.9198
    Precision: 0.3110
    Recall:    0.4908
    AUC:       0.8678
    Best Threshold used: 0.621
Fold 5
  Fold 5 Metrics:
    Accuracy:  0.9371
    Precision: 0.3426
    Recall:    0.3230
    AUC:       0.8478
    Best Threshold used: 0.685

==== Overall Results ====
Mean Accuracy:  0.9282
Mean Precision: 0.3154
Mean Recall:    0.4078
Mean AUC:       0.8562
```

XGBoost

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      6182
           1       0.63      0.53      0.57       329

    accuracy                           0.96      6511
   macro avg       0.80      0.75      0.78      6511
weighted avg       0.96      0.96      0.96      6511
```

Cutoff : 0.4

Accuracy = 95.70

AUC = 0.93

Precision = 0.57

Recall = 0.56

Cutoff : 0.5

Accuracy = 96.07

AUC = 93.71

Precision = 63.37

Recall = 52.58

Cutoff : 0.6

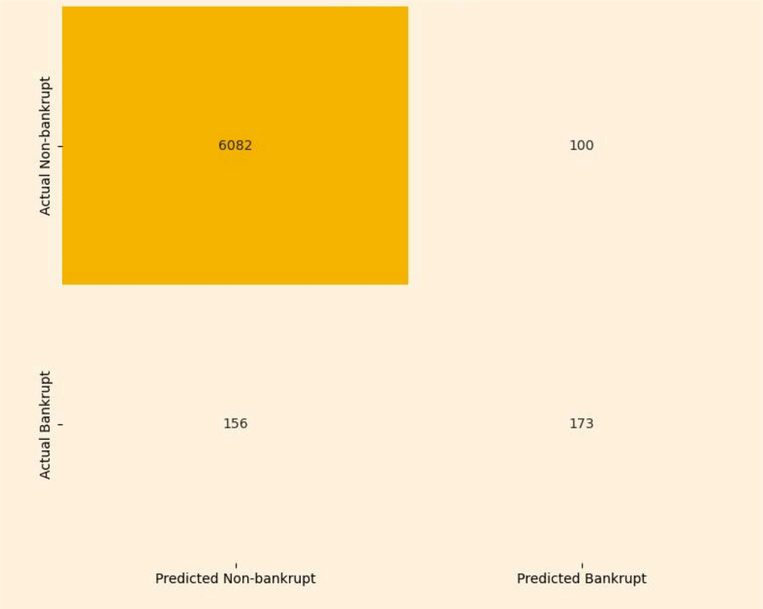
Accuracy = 96.28

AUC = 0.93

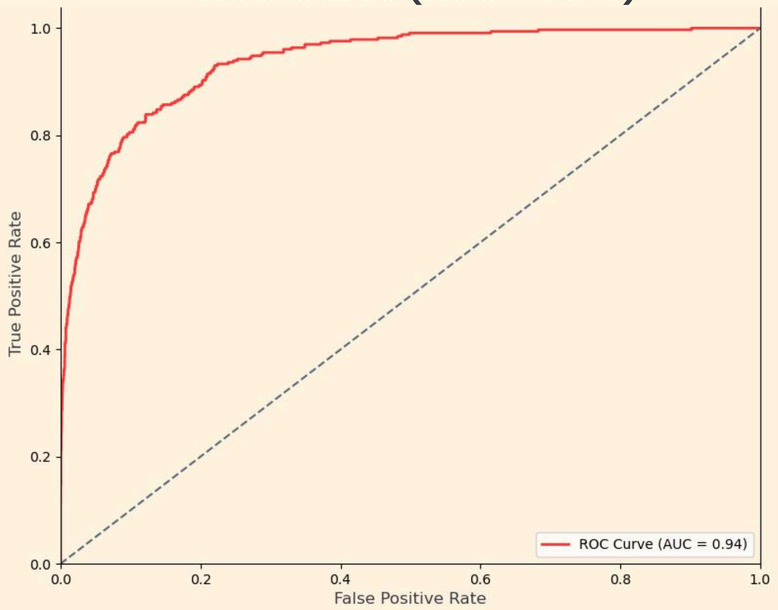
Precision = 0.69

Recall = 0.41

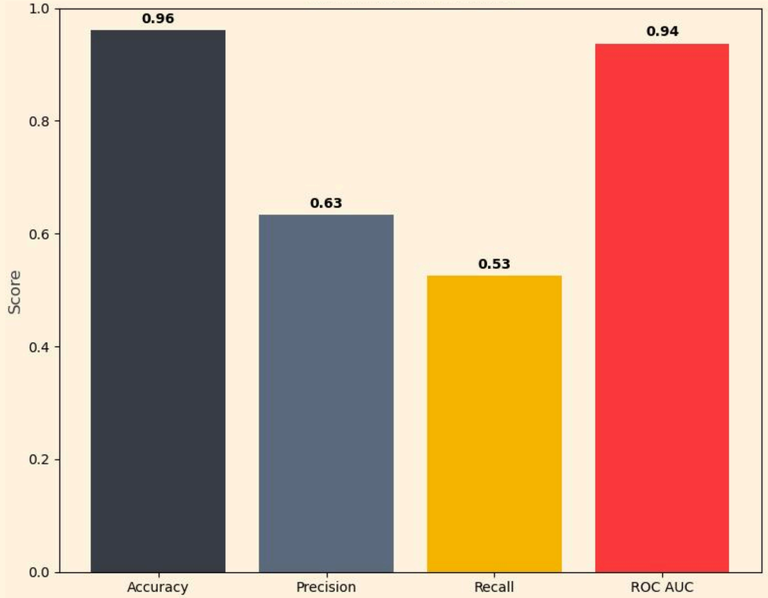
Confusion Matrix



ROC Curve (AUC = 0.94)



Characteristics



Hyper Parameters

```python
# Set XGBoost parameters
params = {
    'max_depth': 11,
    'eta': 0.3,
    'objective': 'binary:logistic',
    'eval_metric': 'auc',

    'min_child_weight': 1,
    'subsample': 0.8,
    'colsample_bytree': 0.8,
    'seed': 42
}
```

```python
# Random Forest hyperparameters
rf_model_fold = RandomForestClassifier(
    n_estimators=100,
    max_depth=10,
    min_samples_leaf=5,
    class_weight='balanced_subsample',
    n_jobs=1,
    random_state=42
)
rf_model_fold.fit(X_train_fold_resampled, y_train_fold_resampled)
```

```python
cat_model = CatBoostClassifier(
    iterations=200,
    learning_rate=0.1,
    depth=8,
    eval_metric='AUC',
    verbose=0,
    random_seed=42
)
```

CatBoost

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      6182
           1       0.63      0.53      0.57       329

    accuracy                           0.96      6511
   macro avg       0.80      0.75      0.78      6511
weighted avg       0.96      0.96      0.96      6511
```

Cutoff : 0.5

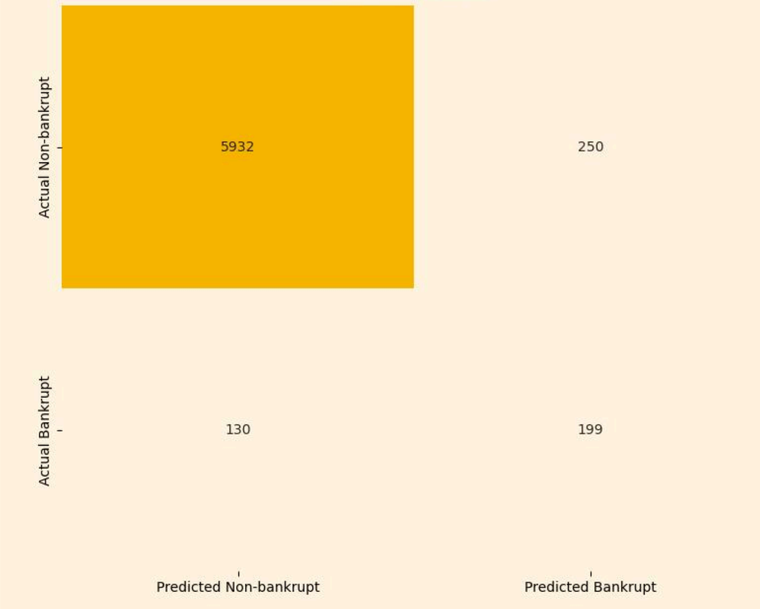
Accuracy = 94.16

AUC = 0.92

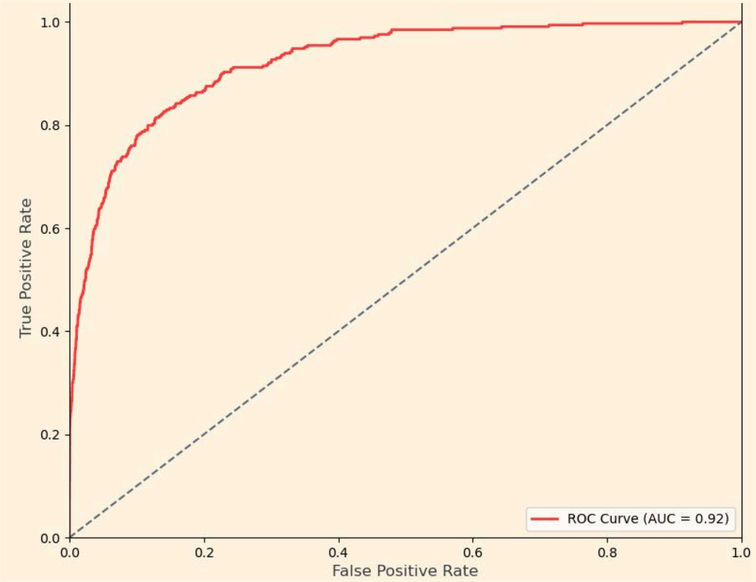
Precision = 0.44

Recall = 0.60

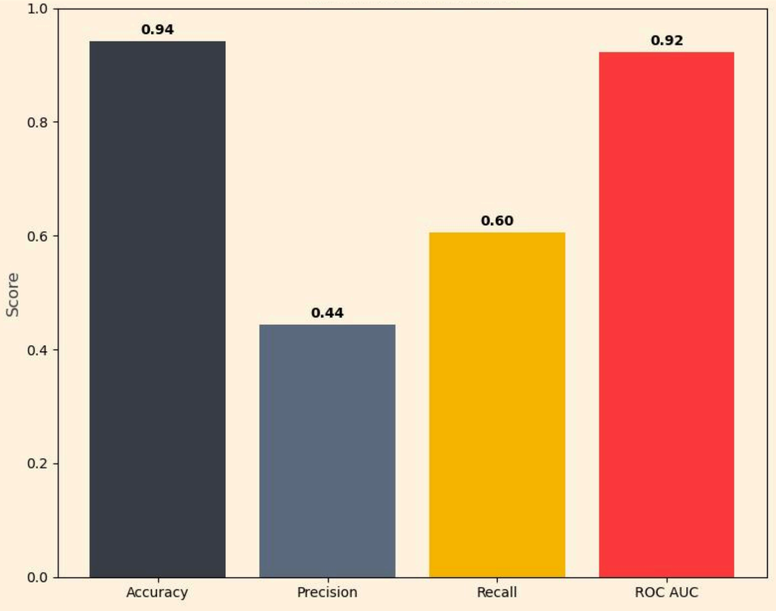
Confusion Matrix



ROC Curve (AUC = 0.92)



Characteristics



Comparision

	AUC	f1 score	precision	recall
Logistic Regression	0.74	0.39	0.10	0.64
Random Forest	0.85	0.33	0.31	0.40
XGBoost	0.93	0.57	0.63	0.52
CatBoost	0.92	0.57	0.44	0.60

Predicted Probabilities

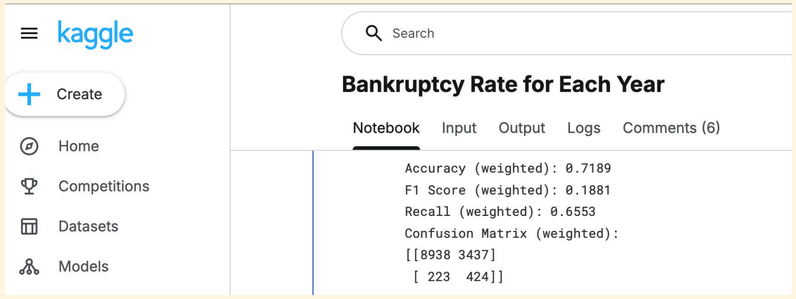
Class 0
Non Bankrupt

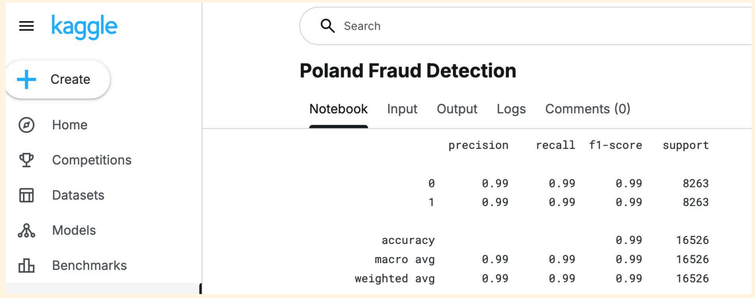
Index	Probability
20310	0.0263
16174	0.0025
22487	0.0034
17440	0.0001
2150	0.0007
1681	0.0027
28999	0.0002
14304	0.0316
24330	0.0001
21077	0.0041

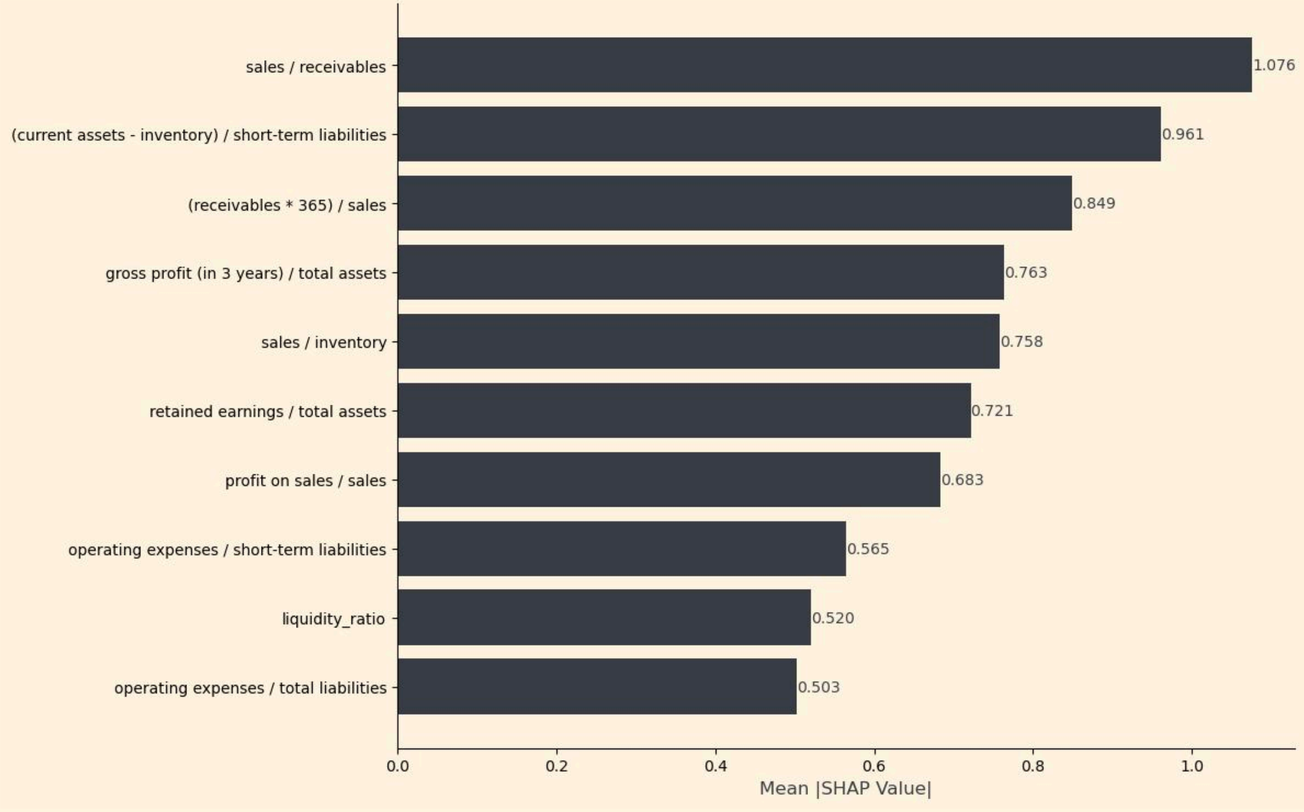
Class 1
Bankrupt

Index	Probability
6793	0.9385
27336	0.8998
27575	0.9885
6757	0.0090
6810	0.1134
6759	0.9957
17141	0.9704
37106	0.8155
6868	0.8793
43005	0.036

Conclusion







Altman Z-Score Formula

$Z=1.2(A)+1.4(B)+3.3(C)+0.6(D)+1.0(E)Z = 1.2(A) + 1.4(B) + 3.3(C) + 0.6(D) + 1.0(E)Z=1.2(A)+1.4(B)+3.3(C)+0.6(D)+1.0(E)$

- A: Working Capital / Total Assets
- B: Retained Earnings / Total Assets
- C: EBIT / Total Assets
- D: Market Value of Equity / Total Liabilities
- E: Total Sales / Total Assets

Key Insight

Our model identified Retained Earnings / Total Assets (B) as one of the most important features.

- This aligns with the Altman Z-Score, reinforcing that our model is focusing on the right financial drivers of bankruptcy prediction.