UNIVERSITÀ
DI PAVIA

CREDIT RISK ANALYSIS

Course: Financial Data Science
Student: F. Amato
Professor: P. Giudici
Accademic year: 2022/23

ABSTRACT

Understand which are the key factors
for a certain level of credit risk to occur

Compare the performance of different ML models
capable to predict the credit risk level for a company
in an year - given past years data

Keywords
Financial Data Science • Credit Risk
Analysis • Machine Learning

GITHUB REPOSITORY

https://github.com/Amatofrancesco99/credit-risk-analysis

TABLE OF CONTENTS

01

INTRODUCTION

WHAT IS CREDIT RISK ANALYSIS

Credit analysis is a type of financial analysis that an investor performs on companies to measure the issuer's ability to meet its debt obligations (calculated using financial ratios, cash flow analysis, trend analysis, and financial projections)

Credit analysis seeks to identify the appropriate level of default risk associated with investing in that particular entity's debt instruments

A review of credit scores and any collateral is also used to calculate the creditworthiness of a business

DATASETS

The main dataset contains the following features (from 2015 to 2020), regarding European companies:

- Name: name of the company
- Turnover: how quickly a company collects cash from accounts receivable or how fast the company sells its inventory
- EBIT: indicator of a company's profitability
- PLTax: company tax based on their cumulative income over their lifetime up until the filing date
- MScore: The higher the M-score level of a company, the more likely the company engages in accounting frauds
- Country: worldwide country of the company

DATASETS

- NACE code: European statistical classification of economic activities
- Sector 1: detailed description of the company's business activities
- Sector 2: general sector
- Leverage: the use of debt to amplify returns from an investment or project
- ROE: measure of a company's annual return
- TAsset: assets owned by the company

Other accessories datasets have been downloaded from Kaggle to extract other useful information, to gain more effective insights

MAIN DATASET

	No	Company name	Turnover.2020	Turnover.2019	Turnover.2018	Turnover.2017	Turnover.2016	Turnover.2015	EBIT.2020	EBIT.2019	...	ROE.2018	ROE.2017	ROE.2016	ROE.2015	TAsset.2020	TAsset.2019	TAsset.2018	TAsset.2017	TAsset.2016	TAsset.2015
0	1	LENDLEASE S.R.L.	29458	16716	9612	8097	7941.0	5600.0	-1556.0	-4540.0	...	8.24	-146.65	60.76	-471.72	49263	28268	15455	15992	13597.0	11659.0
1	2	PRICEWATERHOUSECOOPERS BUSINESS SERVICES SRL (...	16731	16403	16843	12241	9252.0	9515.0	1838.0	841.0	...	61.42	-55.57	-127.29	-87.13	16550	16887	16468	10773	6697.0	8933.0
2	3	EVISO S.P.A.	48568	43039	34302	25791	19760.0	6941.0	1661.0	1464.0	...	57.52	42.73	20.34	44.62	13500	9620	7371	5432	4170.0	2862.0
3	4	CASA SERVICE MACHINE	47999	43484	43043	41682	51267.0	52584.0	416.0	255.0	...	-17.24	0.71	2.89	6.45	24978	25032	25729	21632	25403.0	24941.0
4	5	PANFERTIL SPA	45948	47336	45626	48222	57074.0	62263.0	44.0	713.0	...	-5.17	-6.74	0.03	-8.19	36823	34659	36205	38423	41847.0	41323.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
21249	21250	ASTOR VILLAGE S.R.L.	3161	4635	4742	4499	4277.0	3650.0	985.0	1818.0	...	11.01	8.44	7.83	6.31	15935	15664	14438	13054	12243.0	11695.0
21250	21251	ODONE & SLOA S.R.L.	3161	2562	2559	2334	3692.0	2537.0	60.0	101.0	...	0.62	-4.80	-7.85	-12.84	2487	2317	2351	2521	2797.0	3152.0
21251	21252	GARRIDO MURO SOCIEDAD LIMITADA	3161	3146	2989	3101	2746.0	3154.0	260.0	13.0	...	1.88	3.10	2.90	5.62	2547	1855	1692	1843	1773.0	1699.0
21252	21253	CENTRO INGROSSO JOLLY S.R.L.	3161	2519	2290	2244	1761.0	1821.0	74.0	48.0	...	3.74	3.54	18.85	0.58	2961	2552	2604	2474	1546.0	1222.0
21253	21254	SALONES COMATEL SL	3161	4514	4435	4231	3908.0	2051.0	194.0	733.0	...	15.59	25.60	40.55	-9.98	3576	4259	4747	3993	3027.0	2333.0

121253 rows × 49 columns

02

DATA CLEANSING

COLUMNS STATISTICS

	No	Turnover.2020	Turnover.2019	Turnover.2018	Turnover.2017	Turnover.2016	Turnover.2015	EBIT.2020	EBIT.2019	EBIT.2018
count	121253.000000	121253.000000	121253.000000	121253.000000	121253.000000	121176.000000	1.211080e+05	121249.000000	121252.000000	121252.000000
mean	10110.366259	10857.198313	11571.907903	11147.202164	10545.611812	9864.284776	9.416949e+03	504.857203	586.141820	569.848910
std	5843.476583	9101.352870	9544.166163	9293.686070	8966.629317	8935.990541	9.942035e+03	2086.001182	1797.540232	1949.423069
min	1.000000	2058.000000	2003.000000	2000.000000	2000.000000	0.000000	0.000000e+00	-322920.000000	-139167.000000	-188057.000000
25%	5053.000000	4546.000000	4910.000000	4700.000000	4383.000000	3996.000000	3.689000e+03	60.000000	86.000000	87.000000
50%	10105.000000	7193.000000	7825.000000	7518.000000	7081.000000	6537.000000	6.157000e+03	253.000000	270.000000	265.000000
75%	15157.000000	13680.000000	14753.000000	14208.000000	13401.000000	12463.000000	1.185625e+04	698.000000	717.000000	693.000000
max	21254.000000	49993.000000	49997.000000	49979.000000	49996.000000	294752.000000	1.188225e+06	45155.000000	99633.000000	70371.000000

8 rows × 38 columns

...	ROE.2018	ROE.2017	ROE.2016	ROE.2015	TAsset.2020	TAsset.2019	TAsset.2018	TAsset.2017	TAsset.2016	TAsset.2015
...	121239.000000	121246.000000	121181.000000	121180.000000	1.212530e+05	1.212530e+05	1.212530e+05	1.212530e+05	1.212090e+05	1.212040e+05
...	36.877294	16.076334	16.060269	1.693177	1.301078e+04	1.218059e+04	1.177646e+04	1.131364e+04	1.075000e+04	1.031990e+04
...	4668.108915	1997.679619	1693.639336	5216.355970	3.119684e+04	2.971044e+04	2.922767e+04	2.866877e+04	2.857112e+04	2.859595e+04
...	-369200.000000	-450687.370000	-277108.820000	-1000000.000000	7.100000e+01	5.300000e+01	3.700000e+01	8.600000e+01	0.000000e+00	0.000000e+00
...	3.650000	3.810000	3.380000	2.830000	3.465000e+03	3.174000e+03	2.997000e+03	2.790000e+03	2.517000e+03	2.259000e+03
...	11.350000	11.900000	11.320000	10.660000	6.344000e+03	5.858000e+03	5.579000e+03	5.252000e+03	4.830000e+03	4.474500e+03
...	23.590000	24.830000	24.600000	24.560000	1.284300e+04	1.192900e+04	1.147500e+04	1.087100e+04	1.013900e+04	9.552000e+03
...	1000000.000000	287359.790000	365858.980000	1000000.000000	3.109756e+06	2.597637e+06	2.104548e+06	1.953757e+06	1.993535e+06	2.032843e+06

REMOVE NULL ELEMENTS

```
companies_df.isna().sum().to_frame()
```

	0
No	0
Company name	1
Turnover.2020	0
Turnover.2019	0
Turnover.2018	0
Turnover.2017	0
Turnover.2016	77
Turnover.2015	145
EBIT.2020	4
EBIT.2019	1
EBIT.2018	1
EBIT.2017	1
EBIT.2016	46
EBIT.2015	50
PLTax.2020	2
PLTax.2019	2
PLTax.2018	2
PLTax.2017	2
PLTax.2016	49
PLTax.2015	51
MScore.2020	0
MScore.2019	0
MScore.2018	0
MScore.2017	0
MScore.2016	0
MScore.2015	0
Region	1
Country	0
NACE code	0
Sector 1	0
Sector 2	0
Leverage.2020	0
Leverage.2019	0
Leverage.2018	0
Leverage.2017	0
Leverage.2016	0
Leverage.2015	0
ROE.2020	6
ROE.2019	15
ROE.2018	14
ROE.2017	7
ROE.2016	72
ROE.2015	73
TAsset.2020	0
TAsset.2019	0
TAsset.2018	0
TAsset.2017	0
TAsset.2016	44
TAsset.2015	49

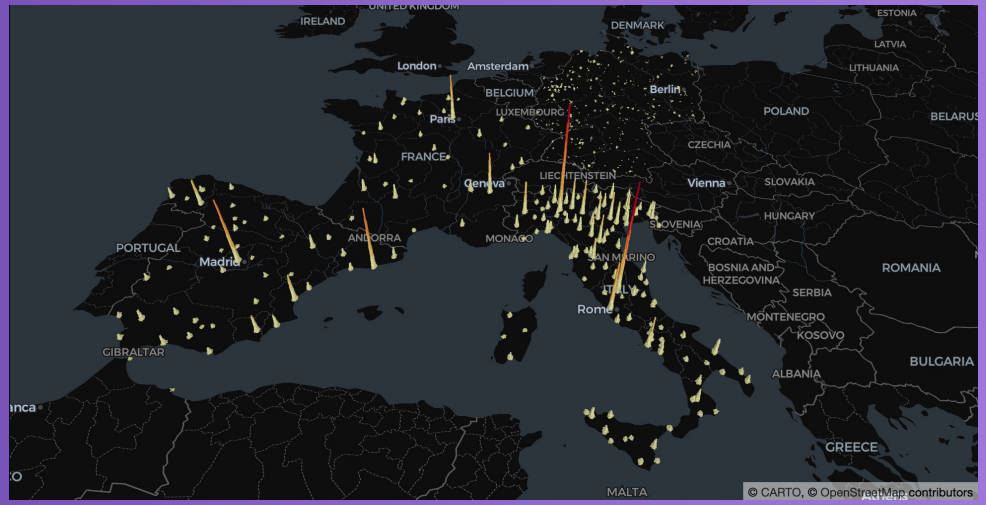
```
original_len = int(len(companies_df))
companies_df = companies_df.dropna()
print('Removed rows:', str(original_len - int(len(companies_df))))
```

```
Removed rows: 245
```

03

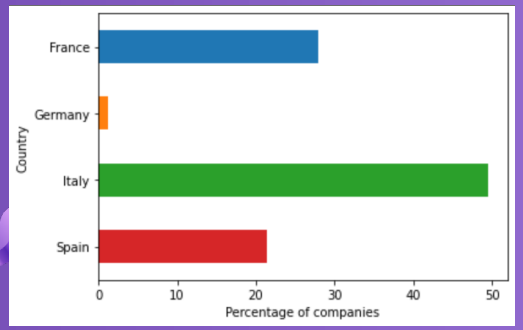
EXPLORATORY ANALYSIS

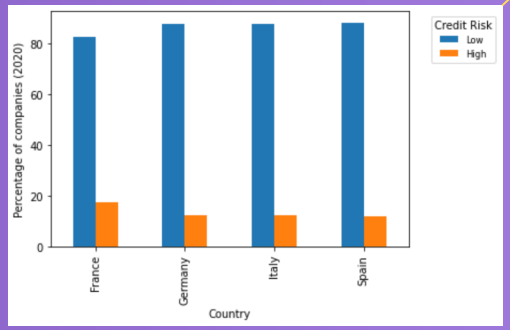
COUNTRIES



Credit risk

- *Low = 0* ⇔ MScore ∈ [A, B]
- *High = 1* ⇔ MScore ∈ [C, D]



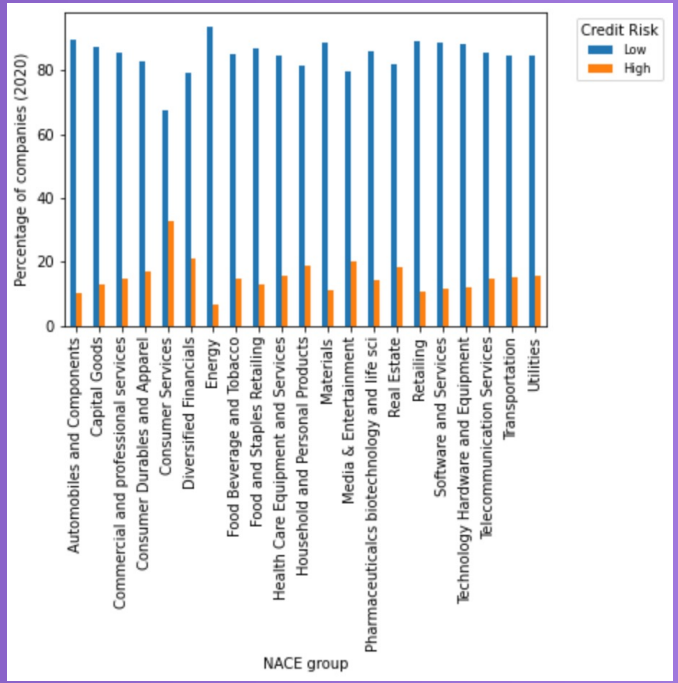


Country	Region	MScore.2020.int 0	1
France	Ain	88.418079	11.581921
	Aisne	82.547170	17.452830
	Allier	83.941606	16.058394
	Alpes-Maritimes	79.387755	20.612245
	Alpes-de-Haute-Provence	96.624473	3.375527
...	...	...	...
Spain	Valencia	91.955307	8.044693
	Valladolid	90.909091	9.090909
	Vizcaya	84.916865	15.083135
	Zamora	95.454545	4.545455
	Zaragoza	88.707654	11.292346

584 rows × 2 columns

SECTORS

Sector 2	%
Automobiles and Components	2.735356
Capital Goods	14.461854
Commercial and professional services	8.953953
Consumer Durables and Apparel	3.950978
Consumer Services	2.948565
Diversified Financials	0.666072
Energy	0.075202
Food Beverage and Tobacco	5.879777
Food and Staples Retailing	7.297038
Health Care Equipment and Services	2.305633
Household and Personal Products	0.348737
Materials	9.881991
Media & Entertainment	2.398189
Pharmaceuticalcs biotechnology and life sci	0.528891
Real Estate	2.004000
Retailing	24.923146
Software and Services	1.832937
Technology Hardware and Equipment	0.866885
Telecommunication Services	0.244612
Transportation	6.339250
Utilities	1.356935



Country	Sector 2	%
France	Automobiles and Components	1.778317
	Capital Goods	14.395195
	Commercial and professional services	13.433542
	Consumer Durables and Apparel	1.600781
	Consumer Services	3.373180
...	...	...
Spain	Software and Services	1.664607
	Technology Hardware and Equipment	0.428704
	Telecommunication Services	0.266492
	Transportation	6.936505
	Utilities	1.602812

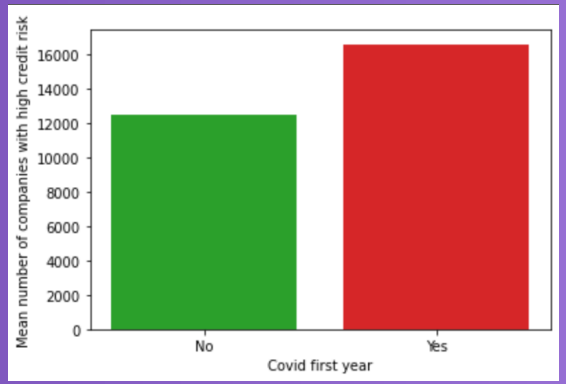
YEARS

Is it possible that after a company gets D credit risk level in an year, the consequent one still gets a D?

```
test_df = companies_df[companies_df['MScore.'+str(year_widget.value)] == companies_df['MScore.'+str(year_widget.value+1)]]
n = test_df[test_df['MScore.'+str(year_widget.value)] == 'D'].count()[0]
```

153 companies had a D credit risk level in 2019 and the consequent year

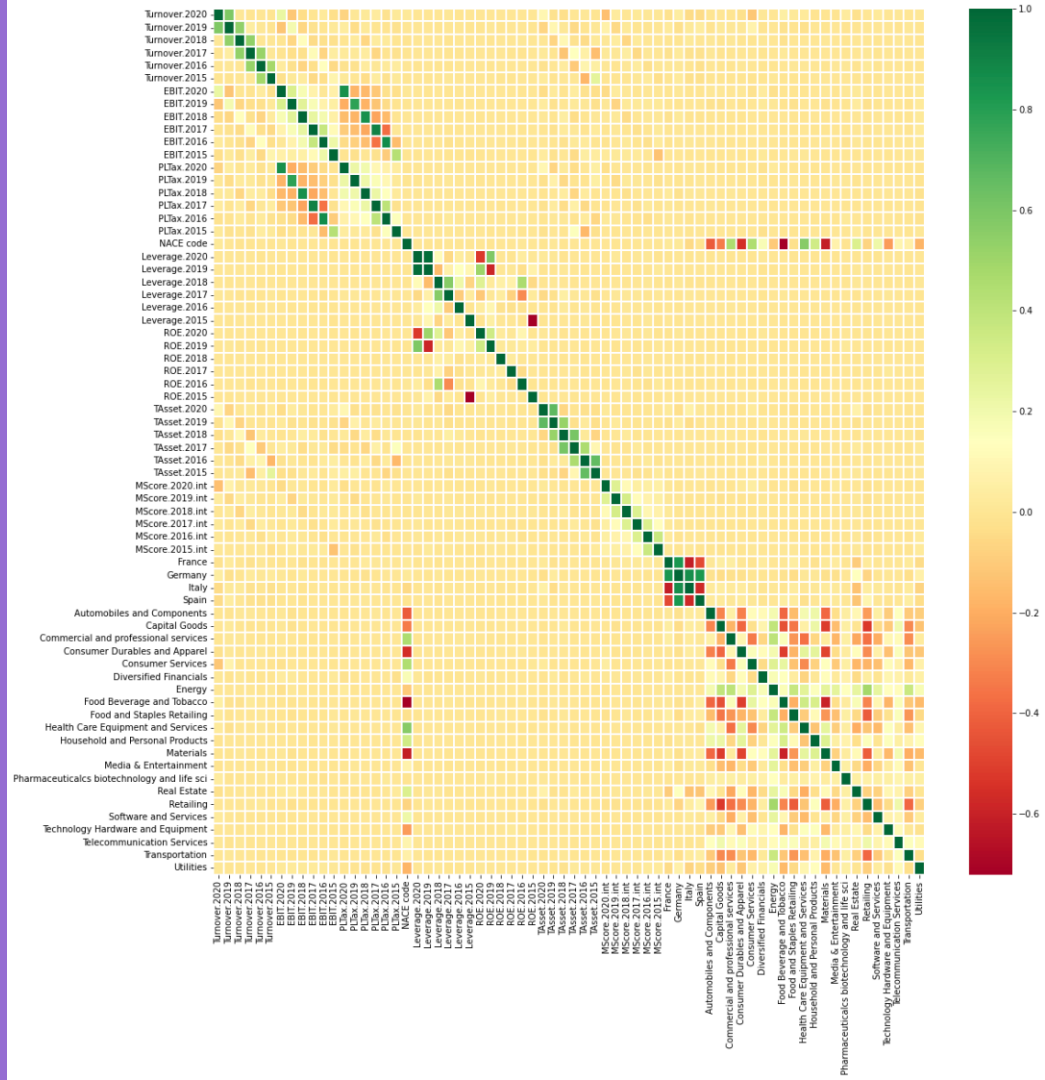
Did Covid-19 led an increase in the number of high credit risk companies?



Percentage increase in high risk level companies in 2020 (with respect to previous years): 32.6%

Correlation is not causation!

PARTIAL CORRELATIONS



04

ML MODELS

UNBALANCED CLASSES

Average number of low level credit risk companies: 107824
Average number of high level credit risk companies: 13184

```
# Fix the unbalanced case
high_risk_df = companies_df[companies_df['MScore.2019.int'] == 1]
low_risk_df = companies_df[companies_df['MScore.2019.int'] == 0].sample(n=len(high_risk_df), random_state=0)
restricted_df = pd.concat([low_risk_df, high_risk_df])
restricted_df.sort_index(inplace=True)
```

Beware of the year choice, because of the already mentioned Covid-19 effect
Objective: predict 2019 credit risk level (selection through slider)

FEATURE SELECTION

Start considering a very easy model (LogReg)
Whether features values are ranging not in [0, 1], MinMax scaling has been applied

Easiest model: credit risk level of previous year, to predict the one of next year

```
X = restricted_df[['MScore.'+str(year_widget.value)+'.int']]
y = restricted_df[['MScore.'+str(year_widget.value + 1)+'.int']]
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0, test_size=0.2, shuffle=True)
X
```

	MScore.2018.int
0	1
4	1
7	0
11	0
16	0
...	...
120994	0
120995	1
120998	0
121003	0
121004	1

```
lr = LogisticRegression(solver='liblinear', random_state=0)
lr.fit(X_train, np.ravel(y_train))
print_performances('Logistic Regression', lr, X_train, y_train, X_test, y_test)
```

```
Logistic Regression
 - Train accuracy: 78.7%
 - Test accuracy: 79.3%

Test              precision    recall  f1-score   support

           0       0.72      0.96      0.83      2467
           1       0.94      0.62      0.75      2397

    accuracy                           0.79      4864
   macro avg       0.83      0.79      0.79      4864
weighted avg       0.83      0.79      0.79      4864
```

FEATURE SELECTION

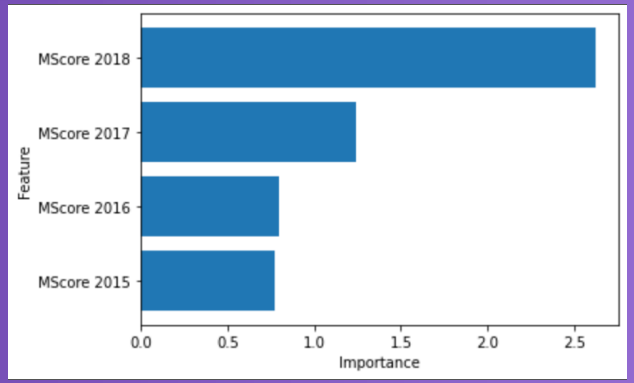
The overall obtained performances can be considered good, but what if we add other features?

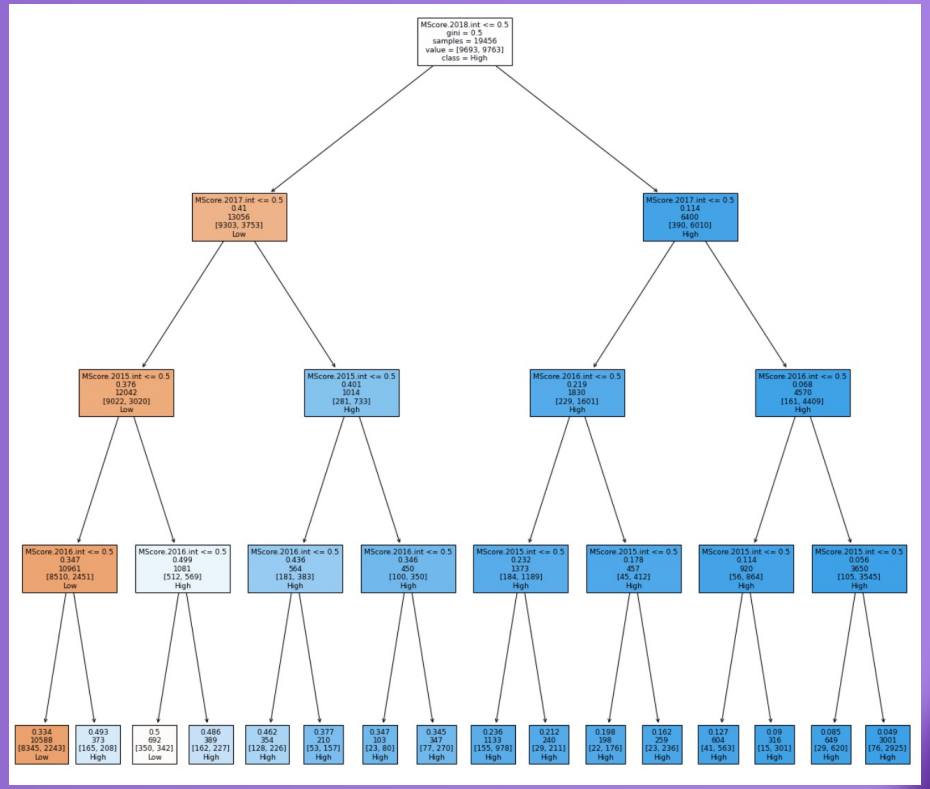
Adding other previous year data like EBIT, Leverage, Turnover, PLTax, ROE, and/or constant features (general sector, country), singularly does not significantly improve the performance metrics. Whereas, combining all of them together it changes, but not so much to justify the loss of interpretability

Add more than one year old data, like all past years credit risk level scores allows to improve significantly the performance metrics (5% of test accuracy increase). Instead, adding other more than one year old data (EBIT, Leverage, ROE, Turnover, ...) does not improve so much the performance metrics

Same occurs when **predicting other years** credit risk levels (features weights and performances metrics change due to **complex variable in time dynamics**)

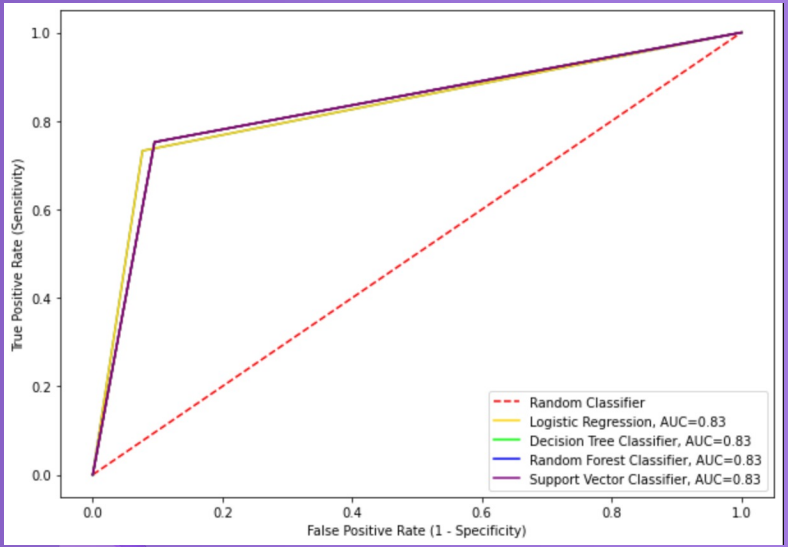
FEATURE IMPORTANCE AND DECISION TREE





MODEL COMPARISON

What happens to performances metrics if we consider other more complex ML models?



THANKS!

QUESTIONS?

Author's GitHub profile

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik