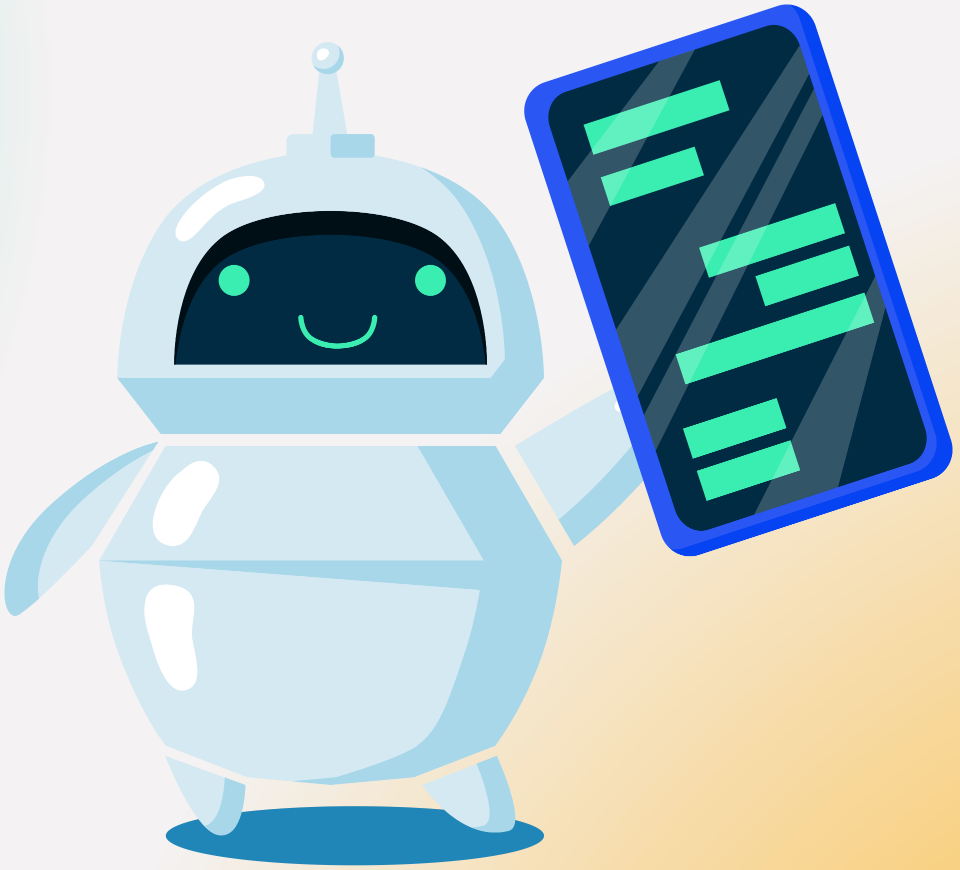
Machine learning for predict the IPO

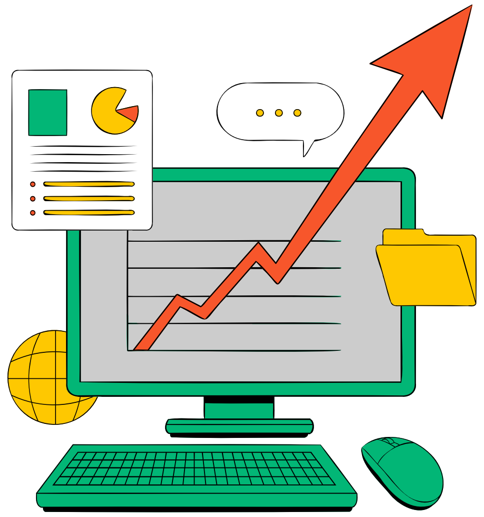
success or failure?

Presentation: Edoardo Pedorcchi

01-The problem
02-The question
03-The dataset
04-The critical point
05-The models used
06-The logistic regression
07-The random forest
08-The comparison
09-The answer



01 - The problem



IPOs are critical for all companies...

will the company be a success or a failure?

will I have made the right or wrong choice?

IPO

02- The question

to be or not to be?

can a machine learning model predict whether a company will flourish or flounder during its IPO?

But more importantly, can it do so using only data from the company's financial statement?

03- The dataset

The dataset contains the analysis of 11 companies(rows).

For each company are reported 8 variables(columns), that are calculated with the companies' financial statement :

- D/E
- EBITDA/revenue
- Net Profit Margin
- current ratio
- Times Interest Earned
- ROA
- ROE
- IPO (only 0=failure and 1=success)

values are only ratios because it does not make sense to use integers values for companies of different sizes and ages.

other 4 companies are used as test dataset

for the data exploration and visualization view the complete project on github!!

04- The critical points

before starting there are 2 problems to be addressed:

small dataset

The data used are few to train a machine learning model

why?

the process to find them is so long...

they will be added over time

few variables:

the training dataset excludes many fundamental variables such as the macroeconomic situation, the stock market situation, the sector in which the company operates, etc.

Why?

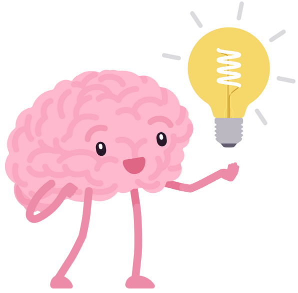
the objective of this model is precisely try to predict IPOs using only financial statement data.

in the future, a model will be created that includes more variables

stay tuned

05 - The models used

Logistic regression



- Advantages:
  - Easy to interpret:
  - Fast training and suitable for small to medium-sized datasets.
  - A good when the relationship is linear or linear in the log transformation.
- Disadvantages:
  - Notfor modeling complex or non-linear
  - Sensitive to outliers in the data.
  - Requires the assumption of linearity

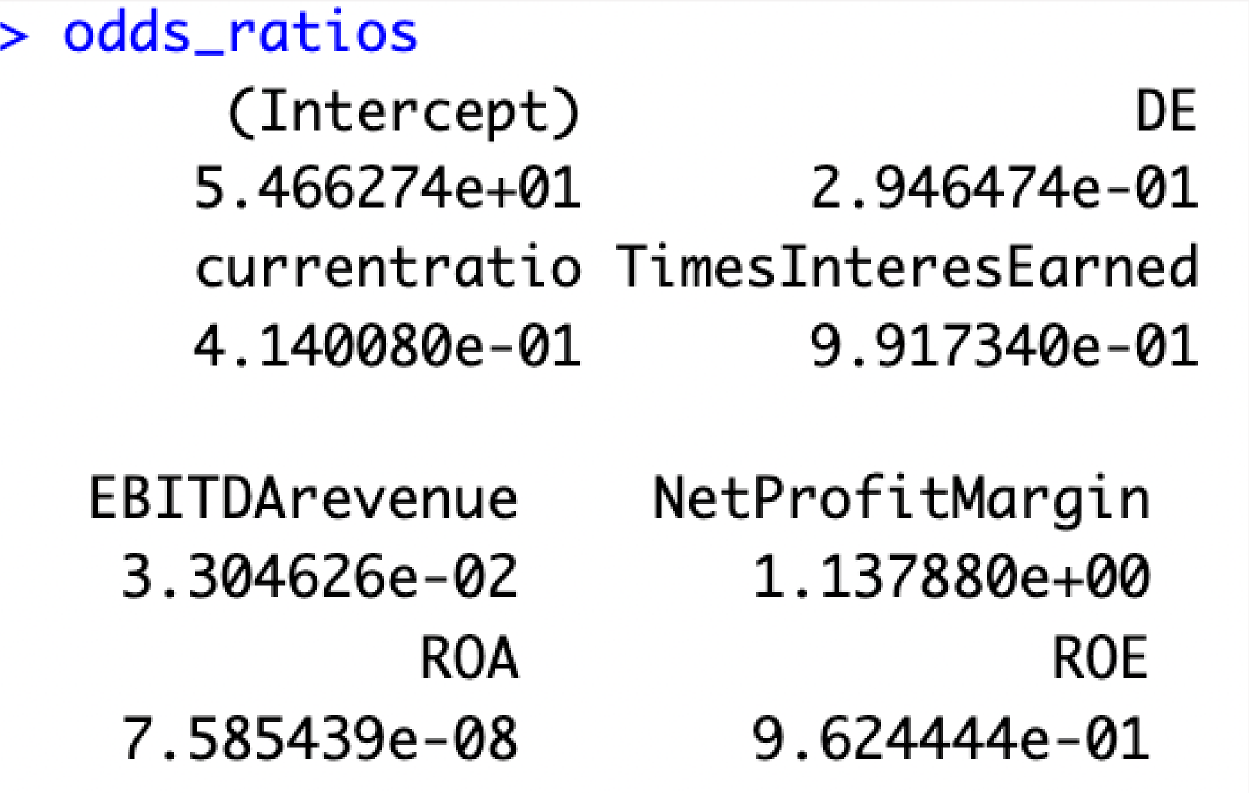
Random forest

- Advantages:
  - Good with complex and non-linear data.
  - Can handle both numerical and categorical data
  - Reduces the risk of overfitting compared to a single decision tree.
  - Provides feature importance(useful for variable selection).
- Disadvantages:
  - require more time for training compared to simpler models (like logistic regression).
  - Less interpretable compared to linear models(like logistic regression.)

06 - The logisitc regression

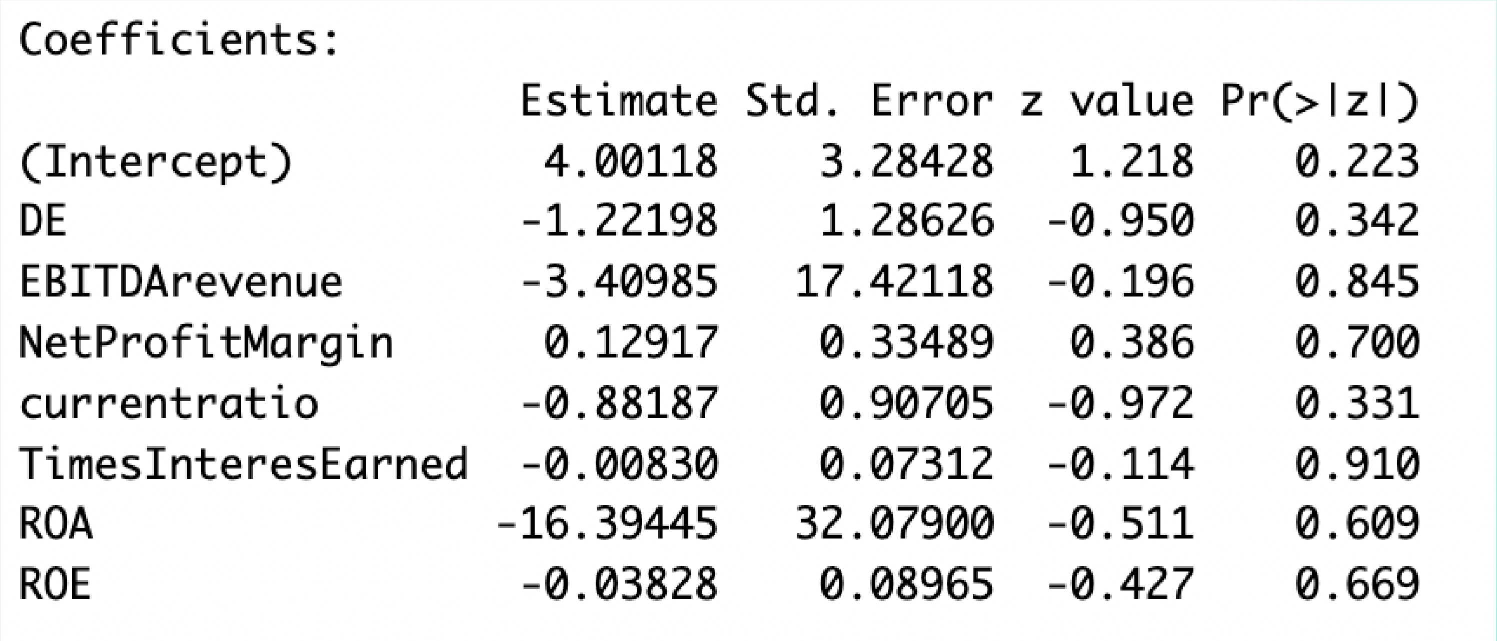
are the variables significant?

view the **odds ratio:**



```
> odds_ratios
      (Intercept)                DE
     5.466274e+01      2.946474e-01
     currentratio TimesInteresEarned
     4.140080e-01      9.917340e-01
 
   EBITDArevenue    NetProfitMargin
    3.304626e-02       1.137880e+00
             ROA                ROE
    7.585439e-08       9.624444e-01
```

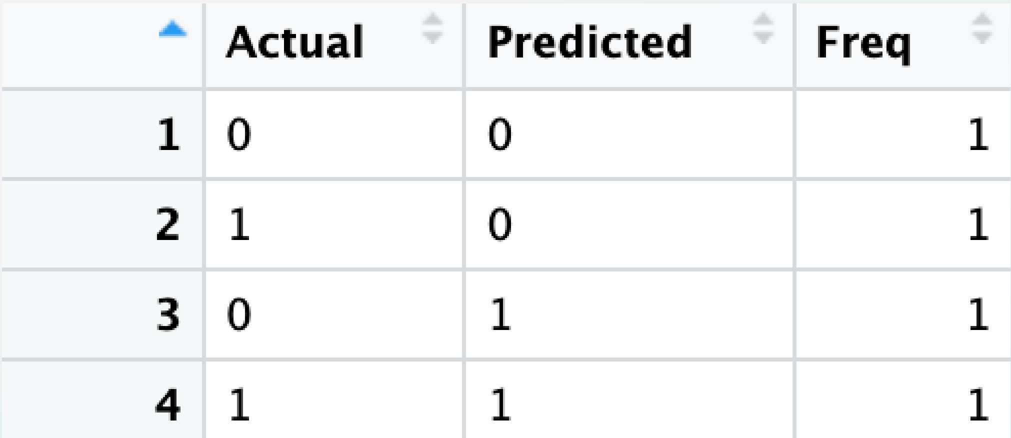
others value



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.00118    3.28428   1.218    0.223
DE                 -1.22198    1.28626  -0.950    0.342
EBITDArevenue      -3.40985   17.42118  -0.196    0.845
NetProfitMargin     0.12917    0.33489   0.386    0.700
currentratio       -0.88187    0.90705  -0.972    0.331
TimesInteresEarned -0.00830    0.07312  -0.114    0.910
ROA               -16.39445   32.07900  -0.511    0.609
ROE                -0.03828    0.08965  -0.427    0.669
```

is the model good?

confusion matrix



	Actual	Predicted	Freq
1	0	0	1
2	1	0	1
3	0	1	1
4	1	1	1

accuracy= 50%

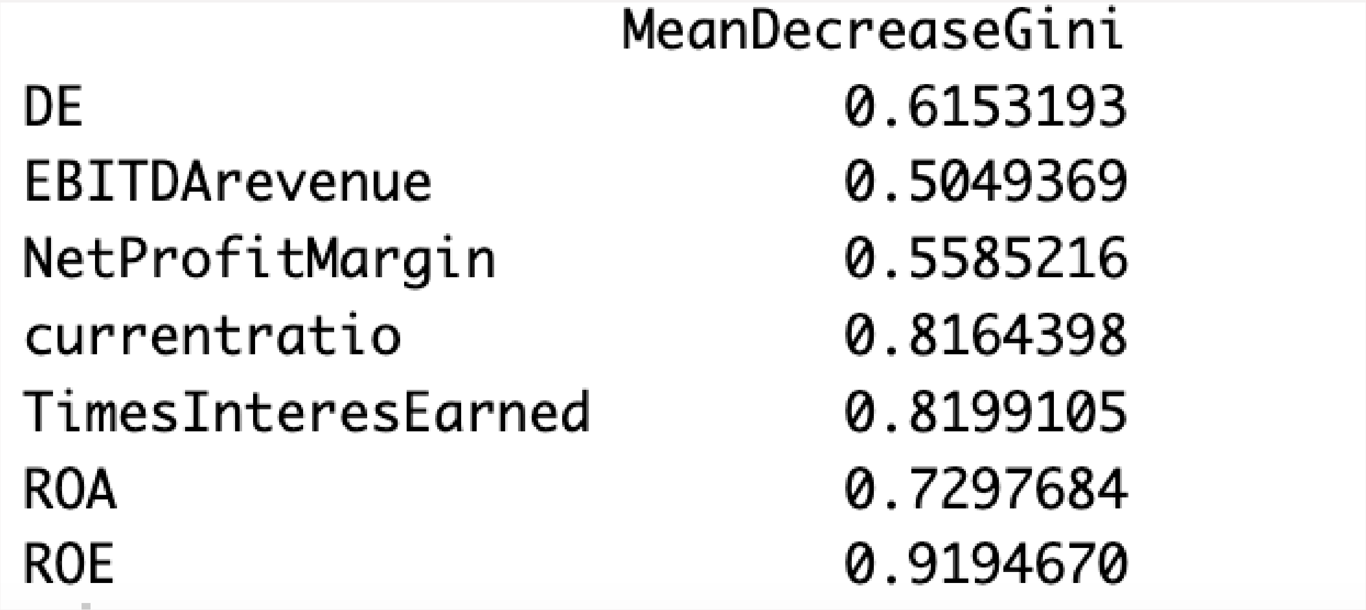
sensitivity= 50%

precision= 50%

07 - The Random forest

are the variables significant?

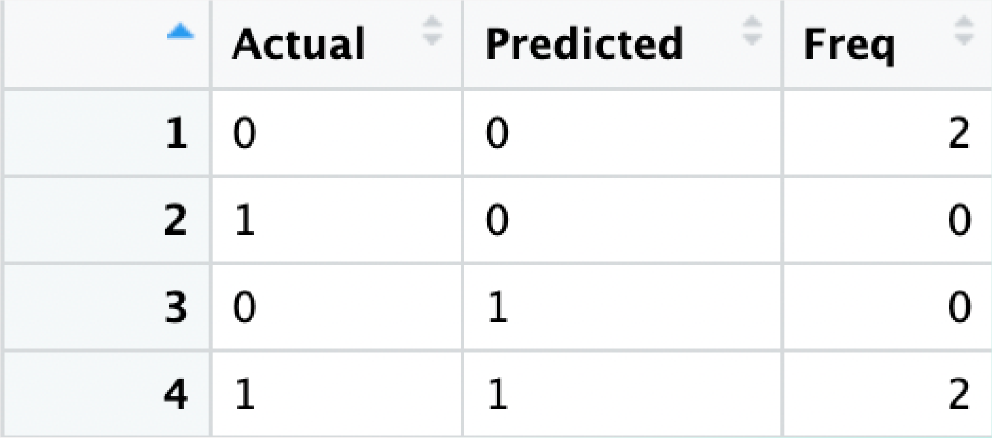
Mean decrease Gini:



```
                   MeanDecreaseGini
DE                        0.6153193
EBITDArevenue             0.5049369
NetProfitMargin           0.5585216
currentratio              0.8164398
TimesInteresEarned        0.8199105
ROA                       0.7297684
ROE                       0.9194670
```

is the model good?

confusion matrix:



	Actual	Predicted	Freq
1	0	0	2
2	1	0	0
3	0	1	0
4	1	1	2

accuracy= 100%
sensitivity=100%
precicion=100%

08 - The Comparison (LR vs RF)

random forest
accuracy=100%

logistic regression
accuracy=50%

such high accuracy is very suspicious

why?

- Lack of Linear Correlation?
- Uninformative Variables?
- few data?

09 - The answer(conclusion)

predicting IPO with ML is possible!!

- random forest work better for this goal
- ROE, TIE, CR are the variables most significative(according to random forest)

more analysis are needed

for example:

- use other models
- add more data
- add more variables
- more trials on treshold

Thanks

for more view :
https://github.com/EdoardoPedrocchi