Corporate Credit Rating Forecasting

Lê Trương Ngọc Liên – DA16

TABLE OF CONTENTS

01

INTRODUCTION TO DATASET

Describe the dataset, the purpose of the model

02

DATA PREPARATION

Explore, analyze, and scale data

03

MODEL BUILDING

Build and select the best model for prediction

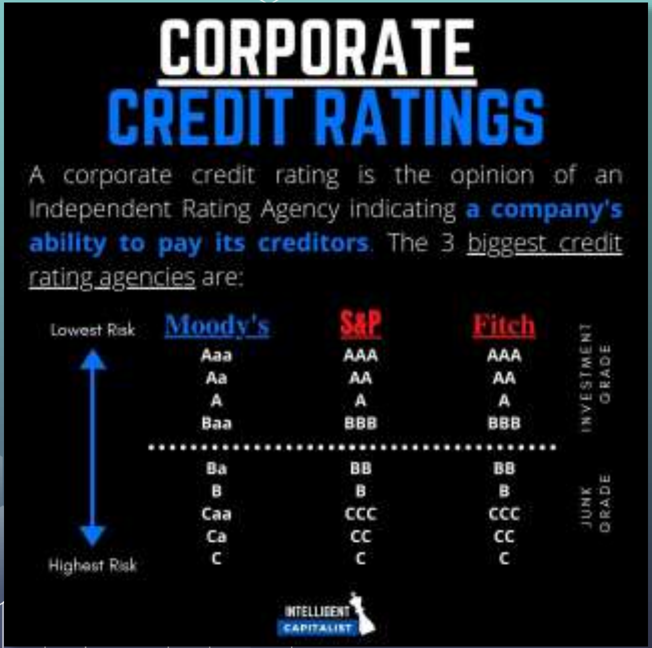
04

PREDICT & EVALUATE

Predict and evaluate the performance of the model

01.

INTRODUCTION TO DATASET



CORPORATE

CREDIT RATINGS

A corporate credit rating is the opinion of an Independent Rating Agency indicating **a company's ability to pay its creditors**. The 3 biggest credit rating agencies are:

	Moody's	S&P	Fitch	
Lowest Risk	Aaa	AAA	AAA	INVESTMENT GRADE
	Aa	AA	AA	
	A	A	A	
	Baa	BBB	BBB	
	Ba	BB	BB	JUNK GRADE
	B	B	B	
	Caa	CCC	CCC	
	Ca	CC	CC	
Highest Risk	C	C	C	

INTELLIGENT CAPITALIST

- Dataset is obtained from Kaggle.

- List of transaction history, business activities, investments, etc. of 593 large US enterprises and companies during the period from 2010 to 2016.

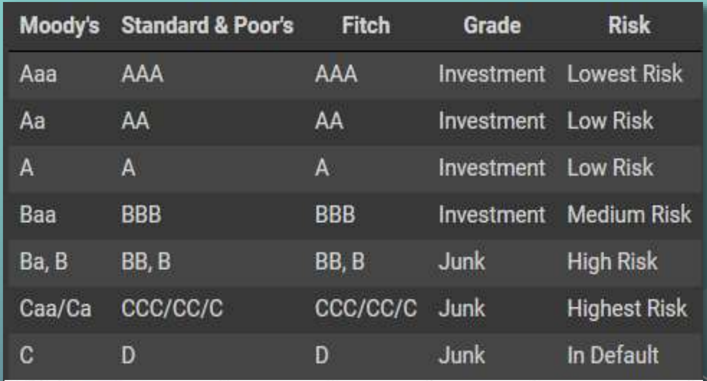
- The dataset has 2029 rows, 31 attributes.

- The purpose of data analysis is to build a machine learning model from credit rating data to predict the rating a company will receive, in order to minimize risks for businesses and investors.

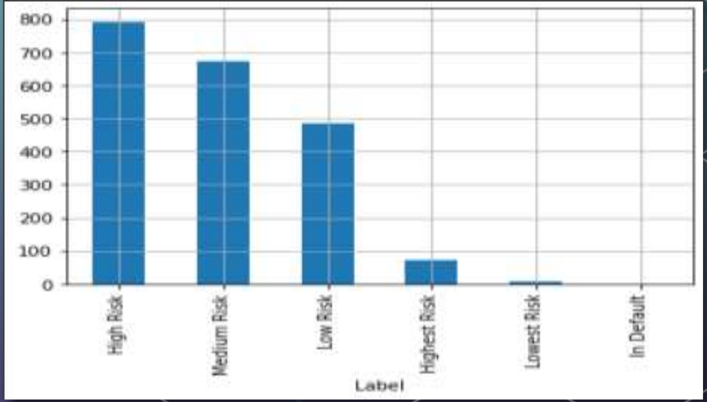
02.
DATA PREPARATION

- In order to create consistency for the rating, as well as reduce the number of classes in the classification process, convenient for the purpose of the assignment, we will replace the rating labels (BBB, BB, A, B, AA, CCC, etc.) AAA, CC, C, D) is equal to the column Rating_name is the meaning of the Ratings. According to the Investopedia website.

- In the Dataset there is only 1 instance with the label "In Default". It is not possible to divide the train or test dataset to build the model. So will delete this instance. Rating_name column has 5 labels

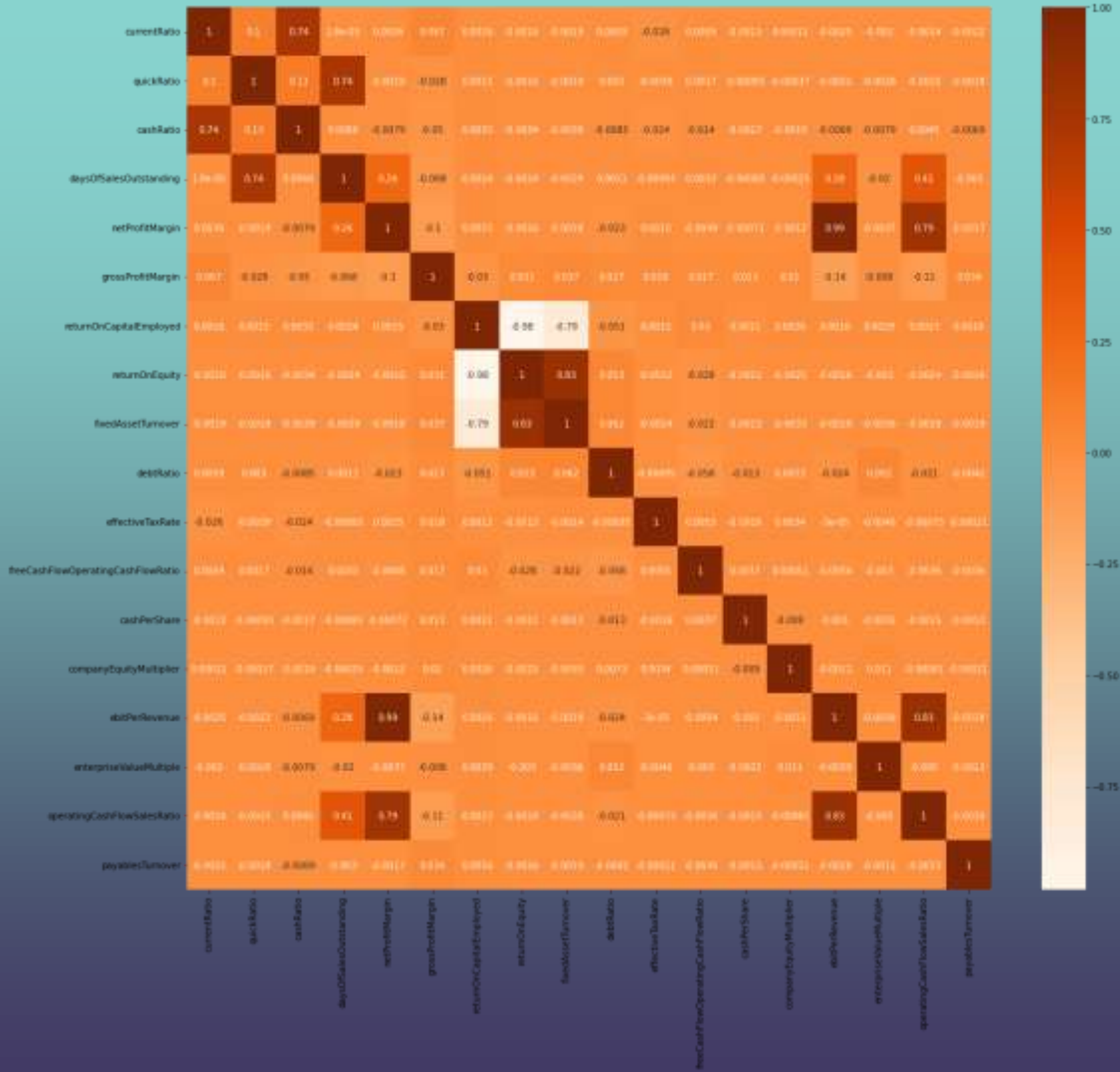


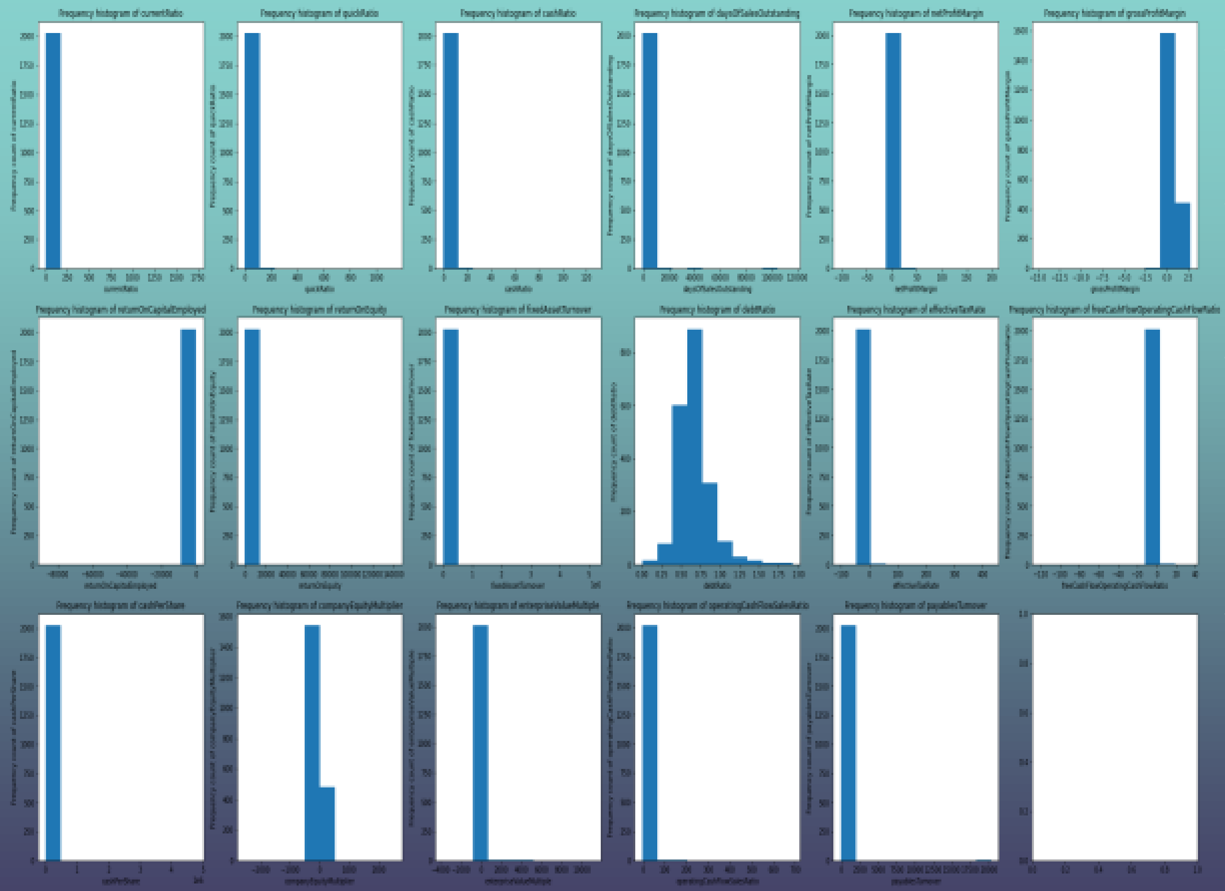
Moody's	Standard & Poor's	Fitch	Grade	Risk
Aaa	AAA	AAA	Investment	Lowest Risk
Aa	AA	AA	Investment	Low Risk
A	A	A	Investment	Low Risk
Baa	BBB	BBB	Investment	Medium Risk
Ba, B	BB, B	BB, B	Junk	High Risk
Caa/Ca	CCC/CC/C	CCC/CC/C	Junk	Highest Risk
C	D	D	Junk	In Default



02.
DATA PREPARATION

- Plot the correlation graph between the variables.
- There are 8 pairs of features that are strongly correlated (> 0.9) with each other. Remove 8 features, keep 17 features for further analysis.





- Most features are not normally distributed.

- The data is heavily affected by outliers.

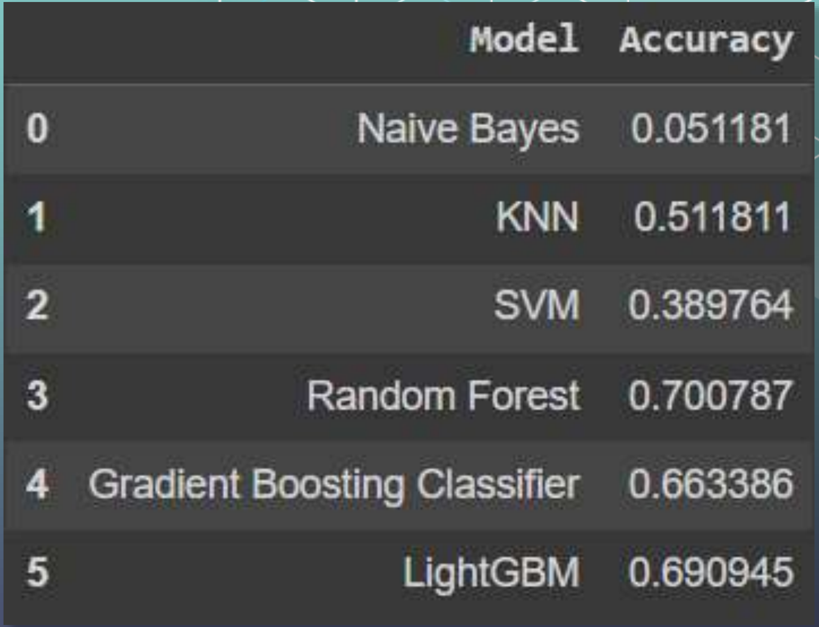
- The distribution of the data could not be commented on. So need to scale the data.

- Since the data is not normally distributed, normalize the data, to return values from any domain to the interval [0,1].

03.
MODEL BUILDING

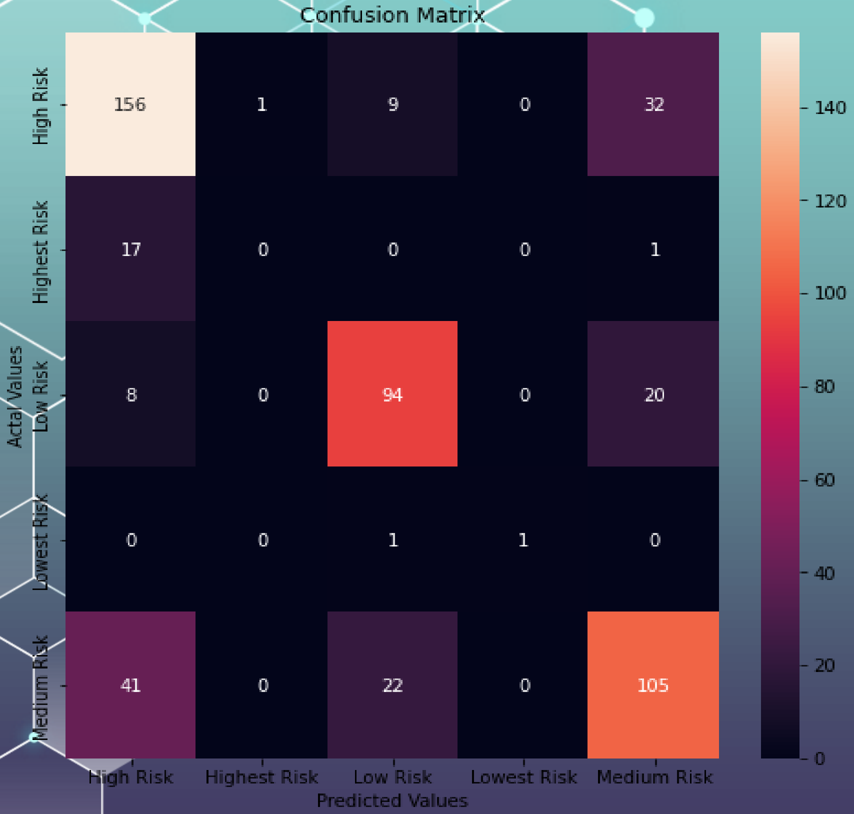
- After processing the data, build 6 models to predict the label.

- Random Forest is the best predictive model, with model accuracy of 70%.



	Model	Accuracy
0	Naive Bayes	0.051181
1	KNN	0.511811
2	SVM	0.389764
3	Random Forest	0.700787
4	Gradient Boosting Classifier	0.663386
5	LightGBM	0.690945

04.
PREDICT & EVALUATE



	High Risk	Highest Risk	Low Risk	Lowest Risk	Medium Risk
High Risk	156	1	9	0	32
Highest Risk	17	0	0	0	1
Low Risk	8	0	94	0	20
Lowest Risk	0	0	1	1	0
Medium Risk	41	0	22	0	105

- **High Risk**

TP = 156
FN = 1 + 9 + 0 + 32 = 42
FP = 17 + 8 + 0 + 41 = 66
TN = 244

- **Highest Risk**

TP = 0
FN = 17 + 0 + 0 + 1 = 18
FP = 1 + 0 + 0 + 0 = 1
TN = 489

- **Low Risk**

TP = 94
FN = 8 + 0 + 0 + 20 = 28
FP = 9 + 0 + 1 + 22 = 32
TN = 354

- **Lowest Risk**

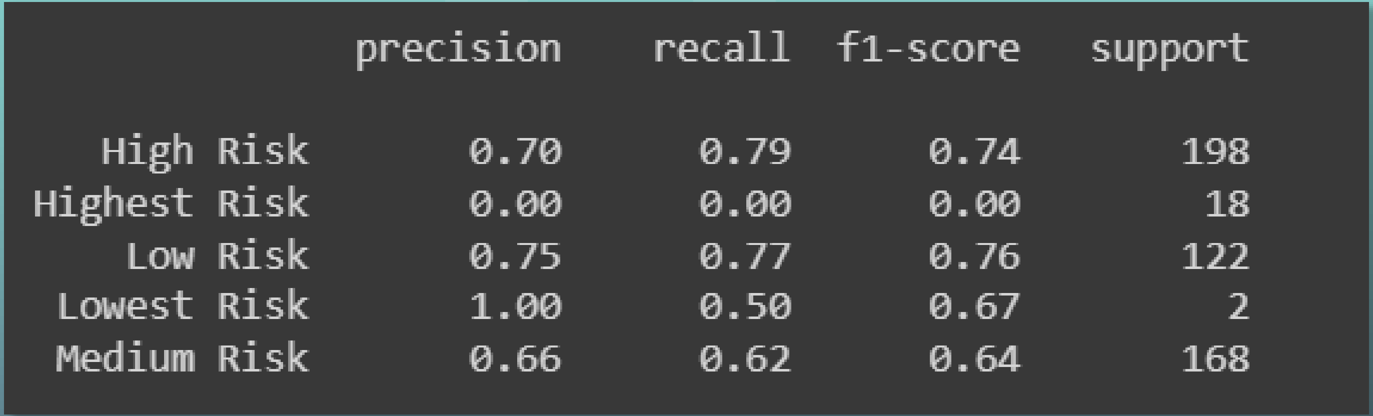
TP = 1
FN = 0 + 0 + 1 + 0 = 1
FP = 0
TN = 506

- **Medium Risk**

TP = 105
FN = 41 + 0 + 22 + 0 = 63
FP = 32 + 1 + 20 + 0 = 53
TN = 287

04.

PREDICT & EVALUATE



```
                precision    recall  f1-score   support

   High Risk       0.70      0.79      0.74       198
Highest Risk       0.00      0.00      0.00        18
    Low Risk       0.75      0.77      0.76       122
 Lowest Risk       1.00      0.50      0.67         2
 Medium Risk       0.66      0.62      0.64       168
```

- The predictive ability of the model for this label is quite good and is highest with F1-Score = 0.74.
- The label "Lowest Risk" has the least data, but the model predicts quite well, with F1 - Score = 0.67.
- In general, most of the labels and models have achieved good classification scores, except for the "Highest Risk" label.

Thank you for listening
Have a good day!!!