

Determinants of Credit Ratings of State-Owned-Enterprises

8,276 - Foundations in Data Science and Machine Learning

Alessandro Fosselard, Antoine Ameye
May 2025



Universität St.Gallen

Very Short Introduction: “What is the current weight of S-O-Es in today’s economy?”

Global Presence

In 2023, SOEs accounted for approximately **12% of global market capitalization**.

Revenue and Assets

Between 2000 and 2023, the number of SOEs among the largest 500 enterprises by revenue worldwide **increased from 34 to 126**.

Regional Variations

In the **OECD area**, the market capitalization of listed firms with more than 25% public sector ownership is **just 2%**. In contrast, this figure is **16% in Latin America** and over **40% in some markets**.

Developing Countries

In developing countries, revenues from businesses with at least 10% state ownership are equivalent to **17% of GDP on average**.

China's SOEs

China's SOEs are particularly prominent, accounting for over 60% of the country's market capitalization and generating **about 23-28% of its GDP**.

Agenda: we want to check whether state-ownership influences the rating, depending on the localization of the state or the industry.

Data Management



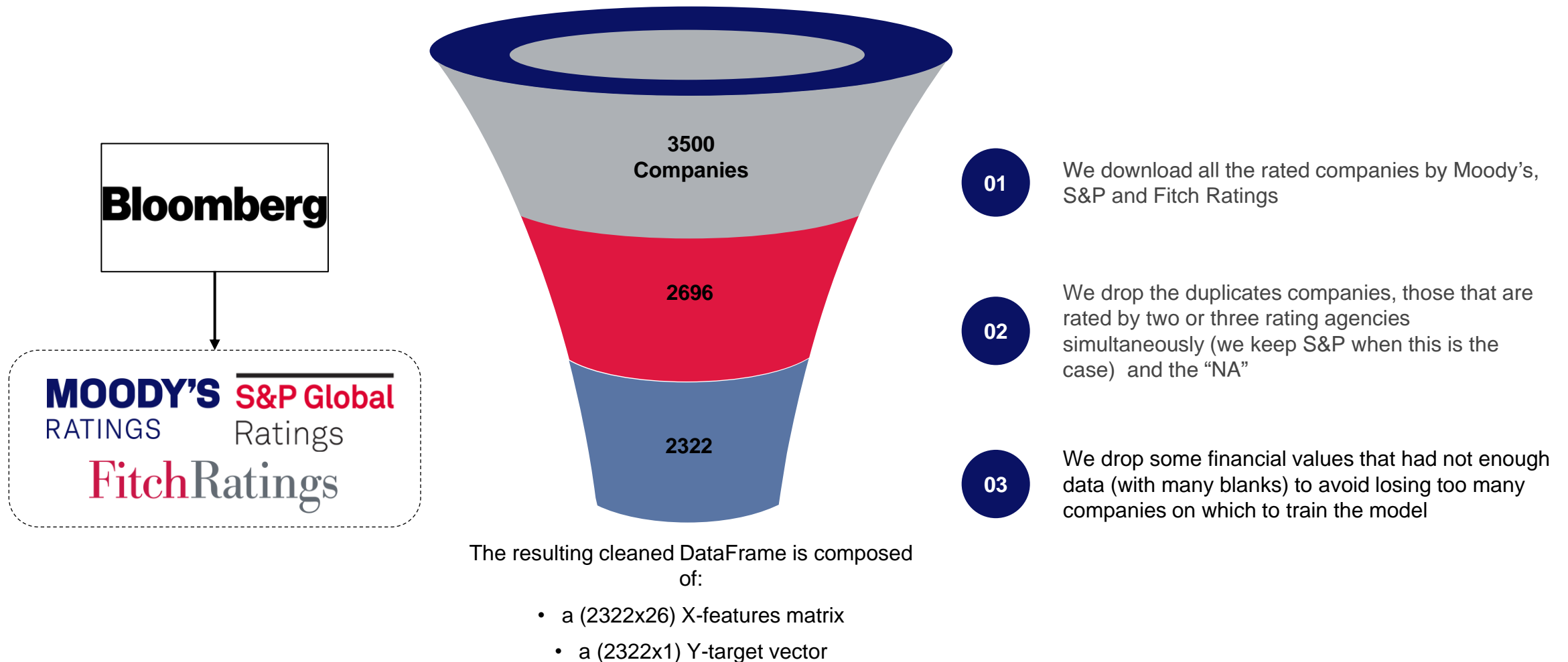
Modeling (ML)



Descriptive Analysis



Getting the data: large DataFrames from Bloomberg® containing 35 columns of data for the 3500 rated firms across the 3 main credit rating agencies.



Data Management

Modeling (ML)

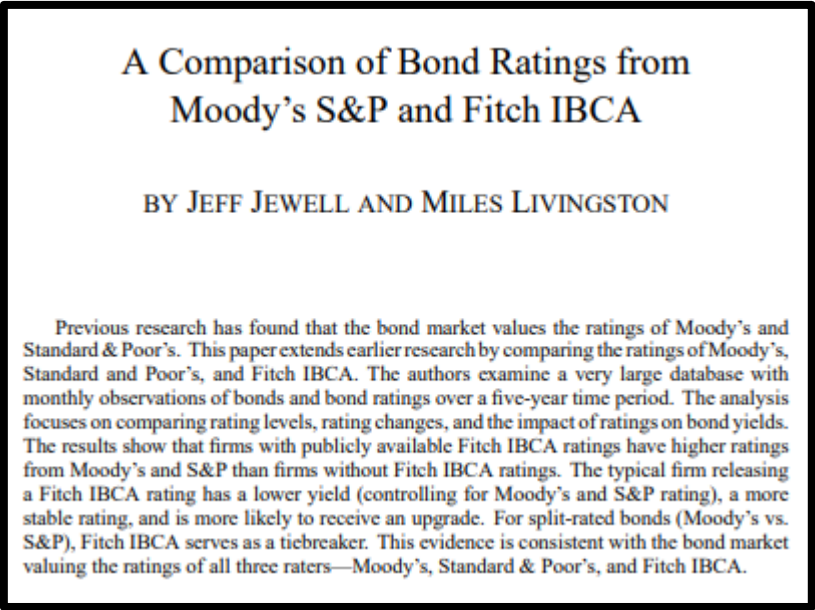
Descriptive Analysis

Main assumption behind the data clustering across the 3 agencies: the rating correlating between the 3 is high and they all have the same number of notches (22).

In 90% of the cases where Moody's, S&P, and Fitch all rated the same bond, Fitch gave the same letter rating as at least one of the other agencies.

Mean rating differences: **Fitch's ratings were, on average, 0.3 notches higher** than Moody's and S&P in the 3-agency sample; in larger samples, the **difference can reach 0.74 notches vs Moody's and 0.56 vs S&P.**

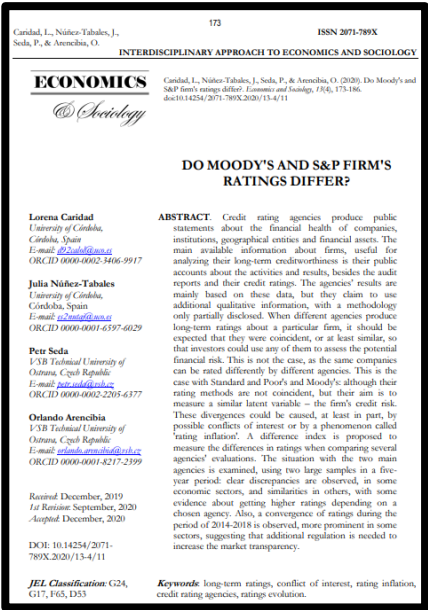
Jewell, J. J., & Livingston, M. (1999). A Comparison of Bond Ratings from Moody's, S&P, and Fitch IBCA.



12–14% of firms had ratings **one notch higher** from S&P.

1–2% had **two or more** notches higher

Caridad, J. M., Arencibia, O., & Seda, P. (2020). Do Moody's and S&P Firm's Ratings Differ?



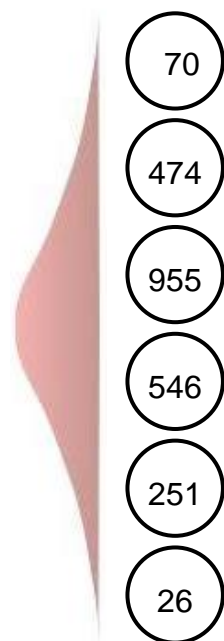
Data Management

Modeling (ML)

Descriptive Analysis

In addition, we solve the problem by regrouping the ratings in 6 categories such that the likelihood of companies being in different groups among credit raters falls drastically.

Distribution of Ratings



S&P	Fitch	Moody's	Grade
AAA	AAA	Aaa	Investment grade: <i>Prime</i>
AA+	AA+	Aa1	<i>High Grade</i>
AA	AA	Aa2	
AA-	AA-	Aa3	
A+	A+	A1	<i>Upper Medium Grade</i>
A	A	A2	
A-	A-	A3	
BBB+	BBB+	Baa1	<i>Lower Medium Grade</i>
BBB	BBB	Baa2	
BBB- (India)*	BBB- (India)*	Baa3 (India)*	
BB+	BB+	Ba1	Non-Investment Grade: <i>Speculative</i>
BB	BB	Ba2	
BB-	BB-	Ba3	
B+	B+	B1	<i>Highly Speculative</i>
B	B	B2	
B-	B-	B3	
CCC+	CCC+	Caa1	<i>Substantial Risk</i>
CCC	CCC	Caa2	
CCC-	CCC-	Caa3	
CC	CC	Ca	<i>Extremely speculative</i>
C	C	C	

Rating Categories

"a High Grade"

"b Upper Medium Grade"

"c Lower Medium Grade"

"d Non-Investment Grade Speculative"

"e Highly Speculative"

"f Substantial Risk"

Observation: we use letter at the beginning of the category title to have an alphabetical order (will be useful for confusion matrix later) and to better define them later on.

Data Management

Modeling (ML)

Descriptive Analysis

Extracting the SOE from the large sample using 47 batches of 50 companies sent to ChatGPT client through API key. Cost: \pm \$2, we lose 5 companies ► 2317

Using OpenAI

Classifying “Yes” or “No”, splitting in 47 batches

- Results: 10.16% of the companies are identified as SOE
- Critic: Quite precise, also catches companies with state minority stakes

Extracting the 236 SOEs from the rest of the companies

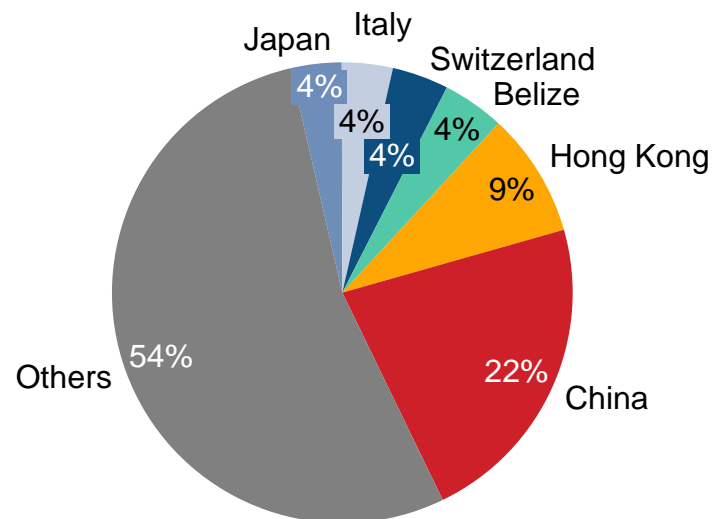
Observation:
*many SOE's are located in China, and the majority of them are in the financial sector, this could result in a model calibrated on the non-SOE firms (which are less skewed towards financials) and **reduce the precision** because the financial ratios of banks are highly influenced by regulations*

Using Keywords

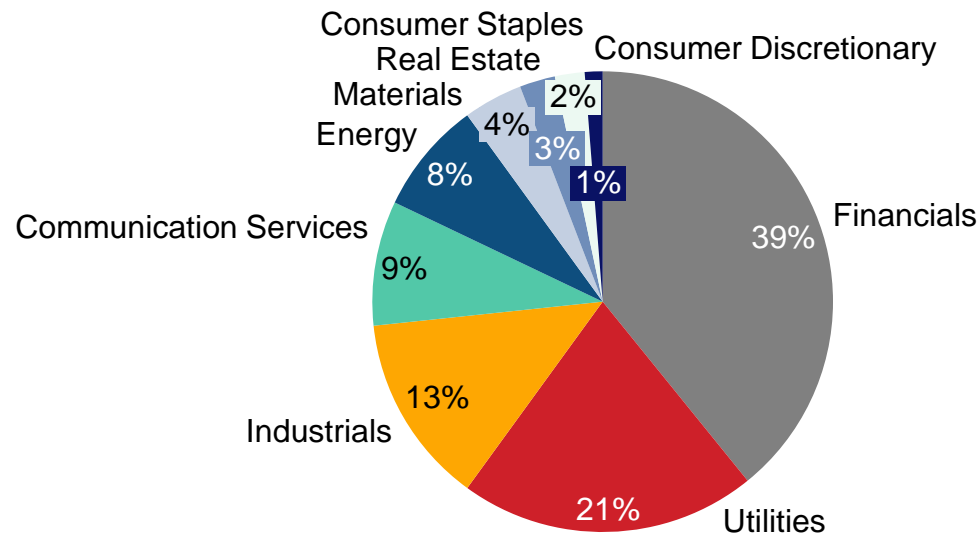
“holding”, “state”, “government” ...

- Results: 14.99% of the companies are identified as SOE
- Critic: Not precise, also many private companies are “holdings”

Where are the SOE's located?



What are the SOE's sectors?

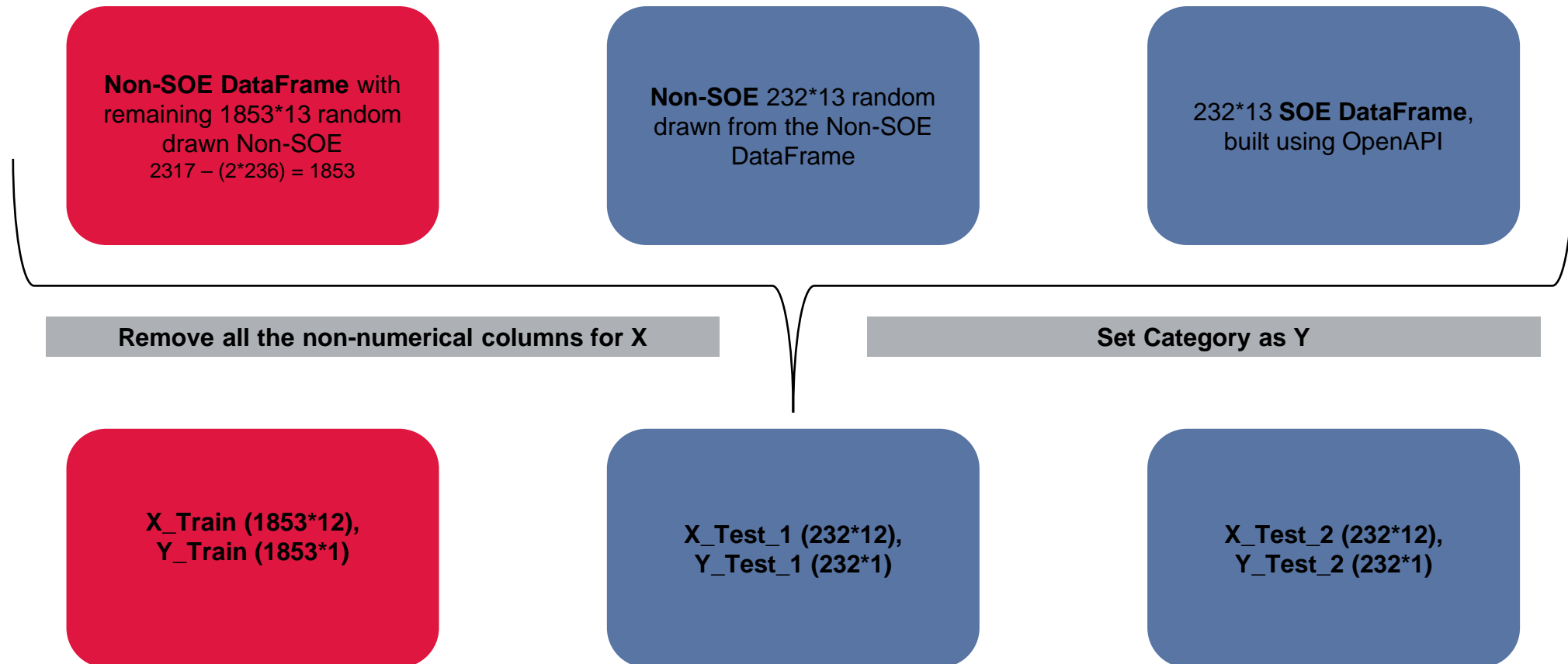


Data Management

Modeling (ML)

Descriptive Analysis

How do we split the data after this?



Data Management

Modeling (ML)

Descriptive Analysis

Model: Logistic Regression with k-Fold Cross-Validation, why did we choose this model?



Model Presentation

Logistic regression is a method that helps us predict categories (like credit ratings) based on input data. It looks at patterns in past data and estimates the probability that something belongs to each possible category.

Linear regression predicts a continuous number (like income or price), while Logistic regression predicts a category (like credit grade A, B, or C) by estimating probabilities for each option.

Objective: Classify **credit ratings** of **non-SOE (state-independent)** firms using **Logistic Regression**.



Code Explanation

We used `LogisticRegressionCV`, a tool that automatically tests the model's performance using internal checks (cross-validation).

Specifically, we applied 5-fold cross-validation, which means that the training data is split into 5 parts. The model is trained on 4 parts and tested on the 5th — and this process is repeated 5 times.

The solver we used, "newton-cg", supports: Multi-class predictions (since we have multiple rating categories).

Data Management

Modeling (ML)

Descriptive Analysis

Comparing two different calibrations of the model

Results with no Penalty

C_value_logit = 1e20

We tested it on a random “virgin”
sample of nongovernment stake

AUC: 79.4939 %
Acc: 52.1186 %

Then we tested it on companies with
government stake

AUC: 64.7098 %
Acc: 38.5593 %

Results with penalty

C_value_logit = [0.01, 0.1, 1, 10, 100]

We tested it on a random “virgin”
sample of nongovernment stake

AUC: 80.5986 %
Acc: 49.1525 %

Then we tested it on companies with
government stake

AUC: 60.8330 %
Acc: 38.1356 %

Data Management

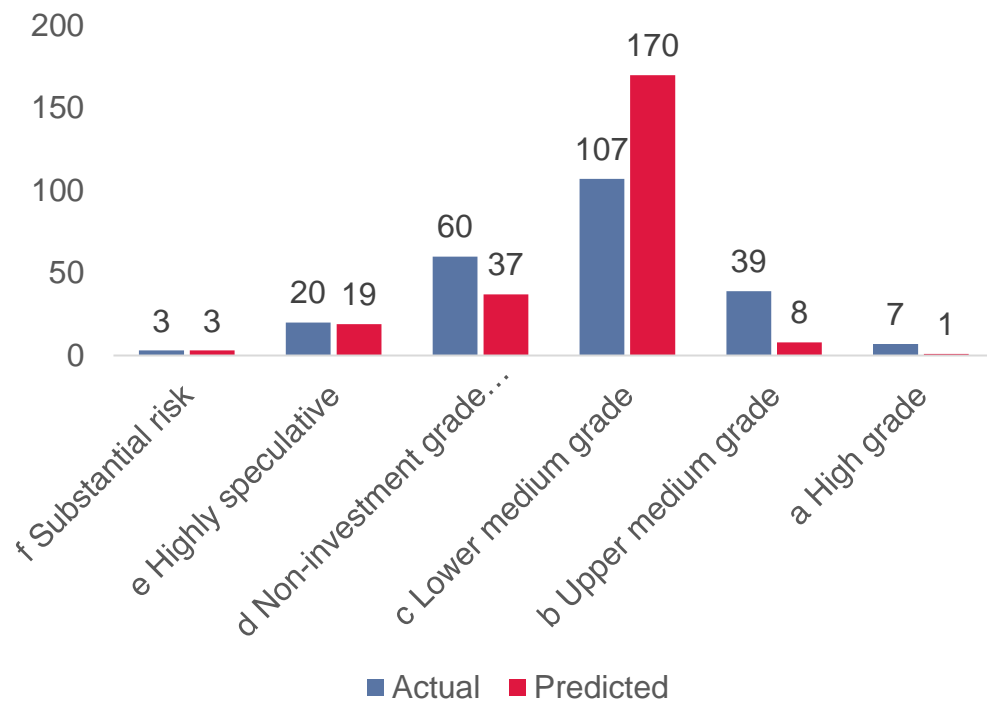
Modeling (ML)

Descriptive Analysis

Logistic Regression - X_TEST_1 (Non-SOE): Centralizing effect, many companies are categorized in the Lower Medium Grade

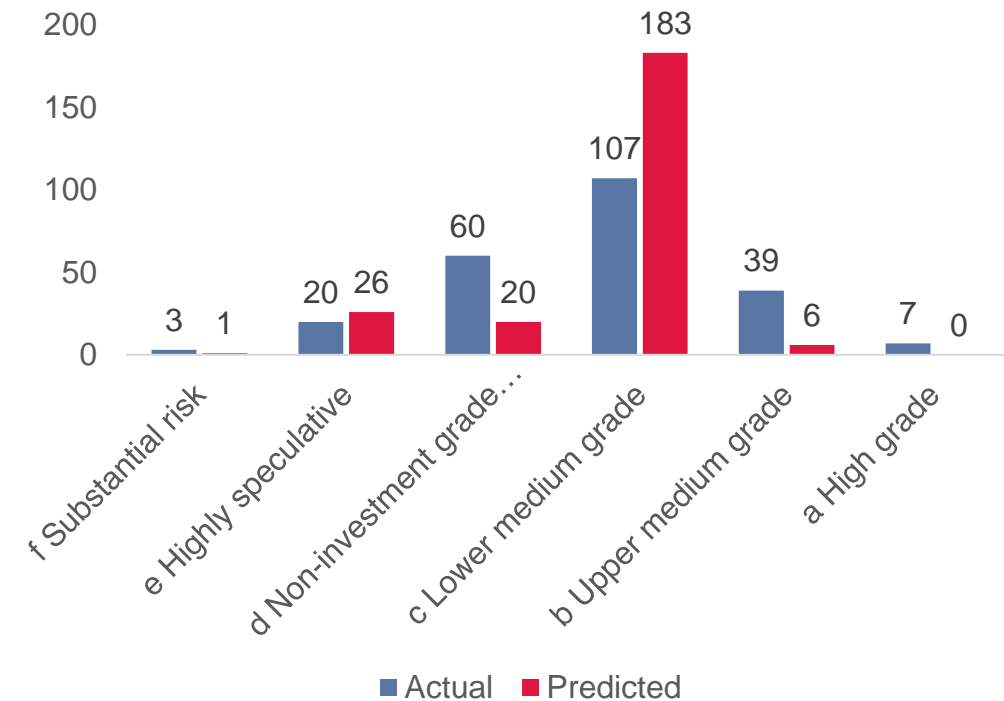
No penalty

The model predicts **less** D and B than the actual number.
The model predicts a lot of C this might be due to the fact that the model was trained on a majority of C.



With penalty

Same as with no penalty but now the model is predicting **more E**.



Data Management

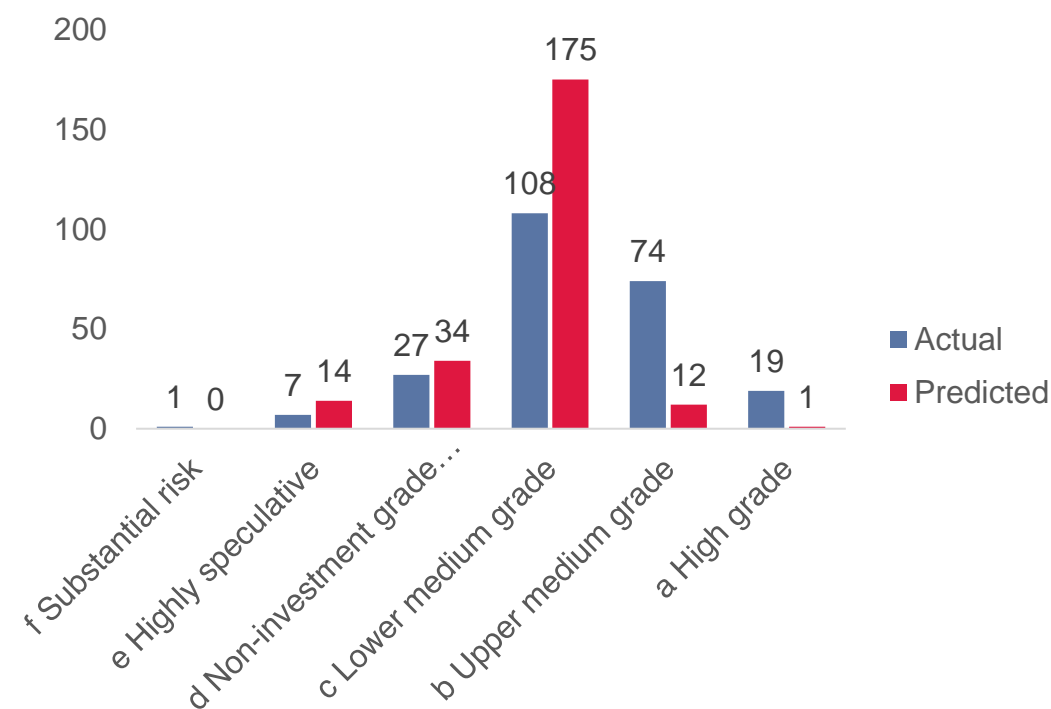
Modeling (ML)

Descriptive Analysis

Logistic Regression - X_TEST_2 (SOE): The model is too negative, which confirms our query, state support seems positive

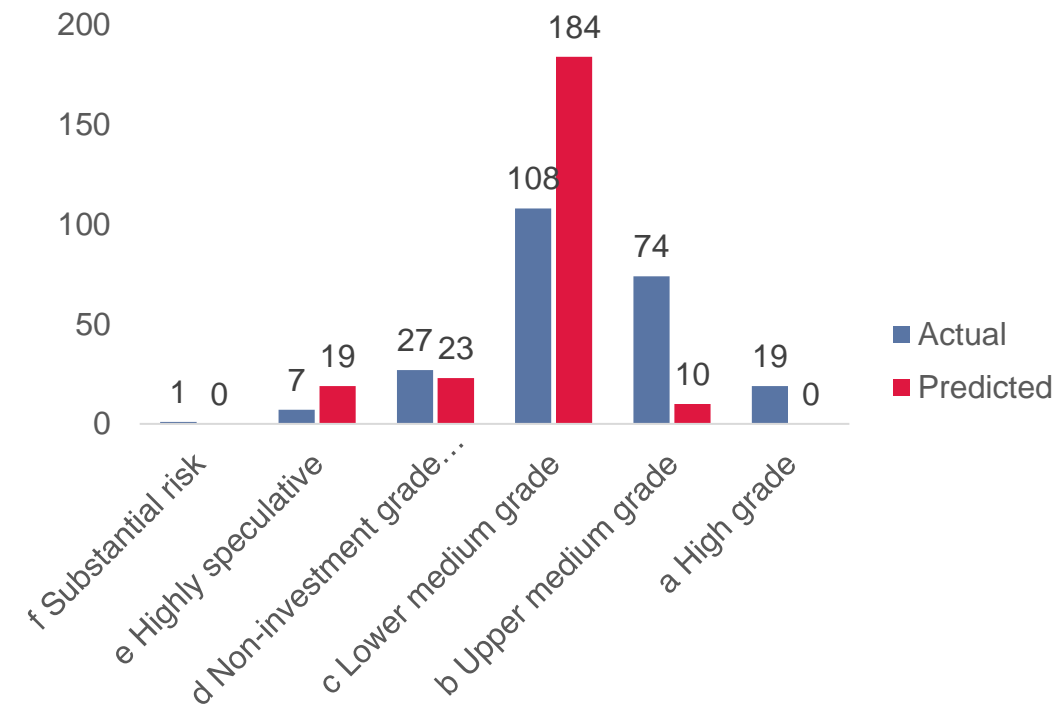
No penalty

This time the model predicts **more** D and E than there actually are. However, it still predicts **less** A and B, but now this is consistent with a potential state benefit.

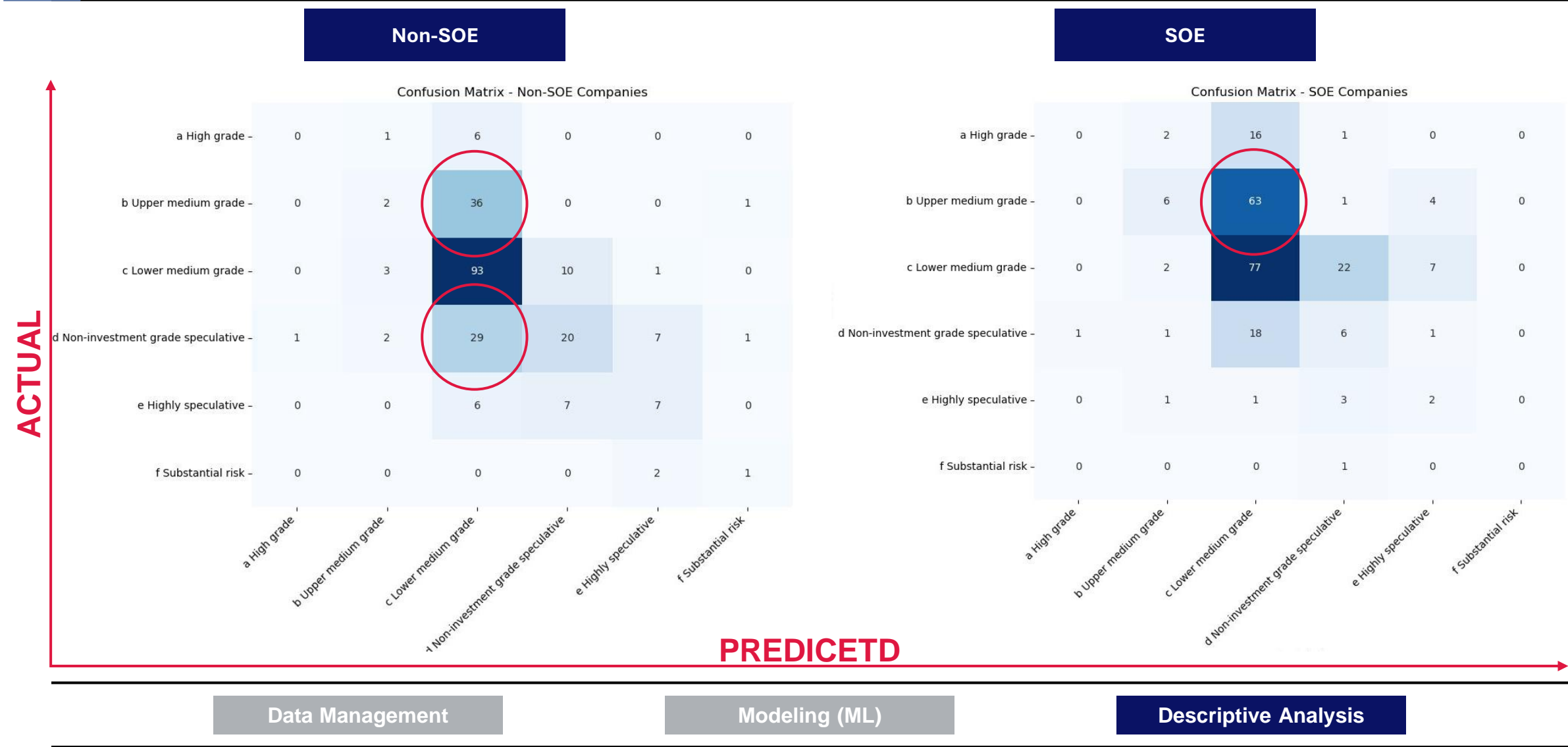


With penalty

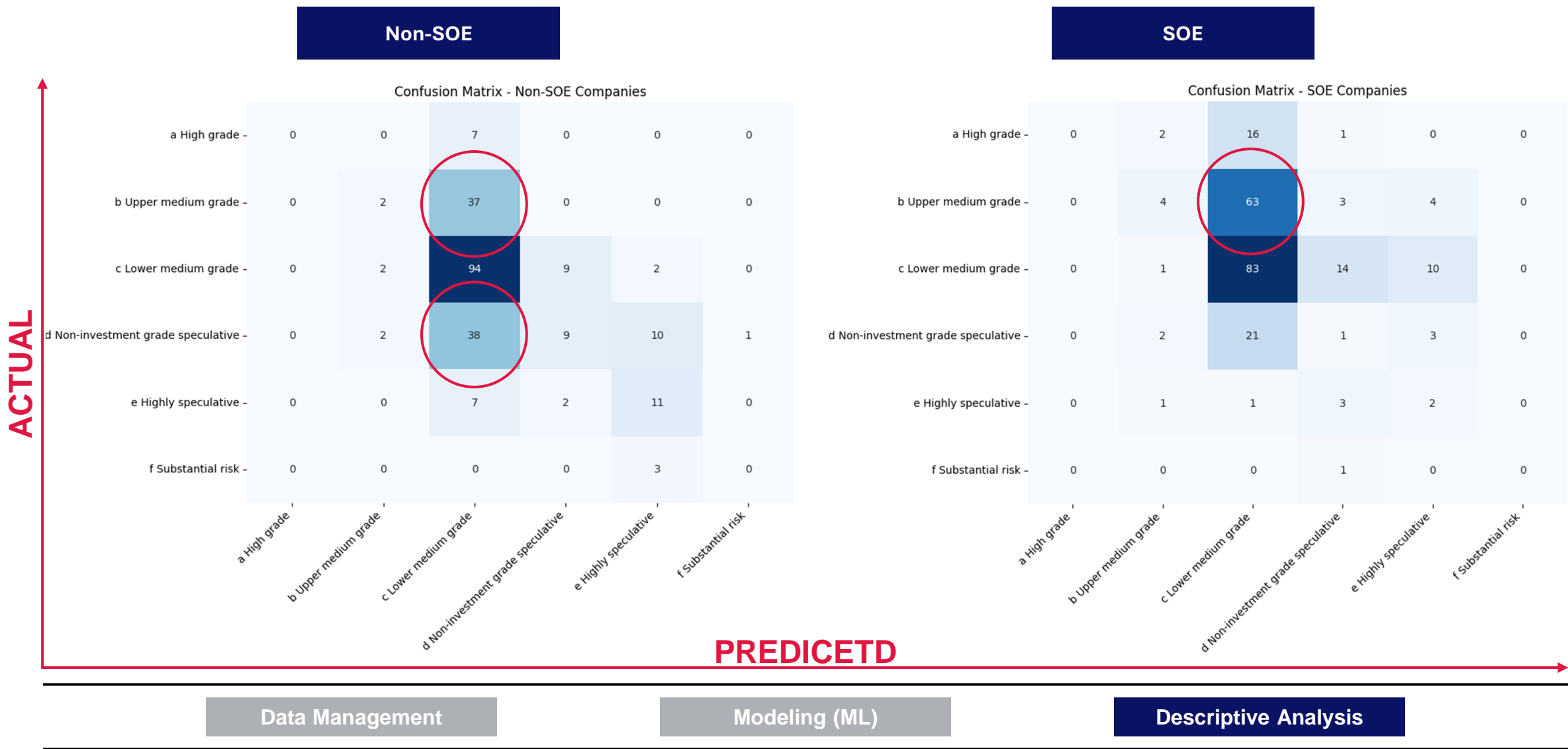
The model predicts **less** D but **more** E than the actual number. This is also in line with a potential state benefit. It still predicts **less** A and B than the actual number.



Confusion Matrix Logistic Regression: in the SOE a lot of mistakes were from C to B. Without Penalty



Confusion Matrix Logistic Regression: in the SOE a lot of mistakes were from C to B. With Penalty



Model: Random Forest Model



Model Presentation

Random forests are built from decision trees:
“**Trees have one aspect that prevents them from being the ideal tool for predictive learning, namely accuracy**” (The Elements of Statistical Learning)

Random Forests are **not flexible** when it comes to classifying new samples because decision rules are hard-coded during training meaning they only work well when new data looks similar to the training data. It **struggles to adapt** to new, unseen patterns.



Why it might be better?

Credit ratings are based on **rule-like decisions**
Rating agencies use structured scorecards: if leverage is high, downgrade; if interest coverage is strong, upgrade.
Random Forests are well suited to replicate this rule-based logic.

Nonlinear interactions matter in credit: A company with high debt but massive cash flow is very different from a company with the same debt and no cash. **Random Forests naturally handle nonlinear combinations** of variables without you having to specify interactions manually.

Data Management

Modeling (ML)

Descriptive Analysis

Random Forest: Results



Results

We tested it on a random “virgin” sample of nongovernment stake

AUC: 88.25%
Acc: 57.20%

Then we tested it on companies with government stake

AUC: 72.28%
Acc: 43.22%



Description of the code

We used a **pipeline**, meaning all steps (data preparation + model) were packaged together for cleaner, repeatable analysis.

We used **class_weight='balanced'**, which automatically adjusts the importance of each class based on how frequent it is in the training data. This useful because our rating classes were not evenly distributed.

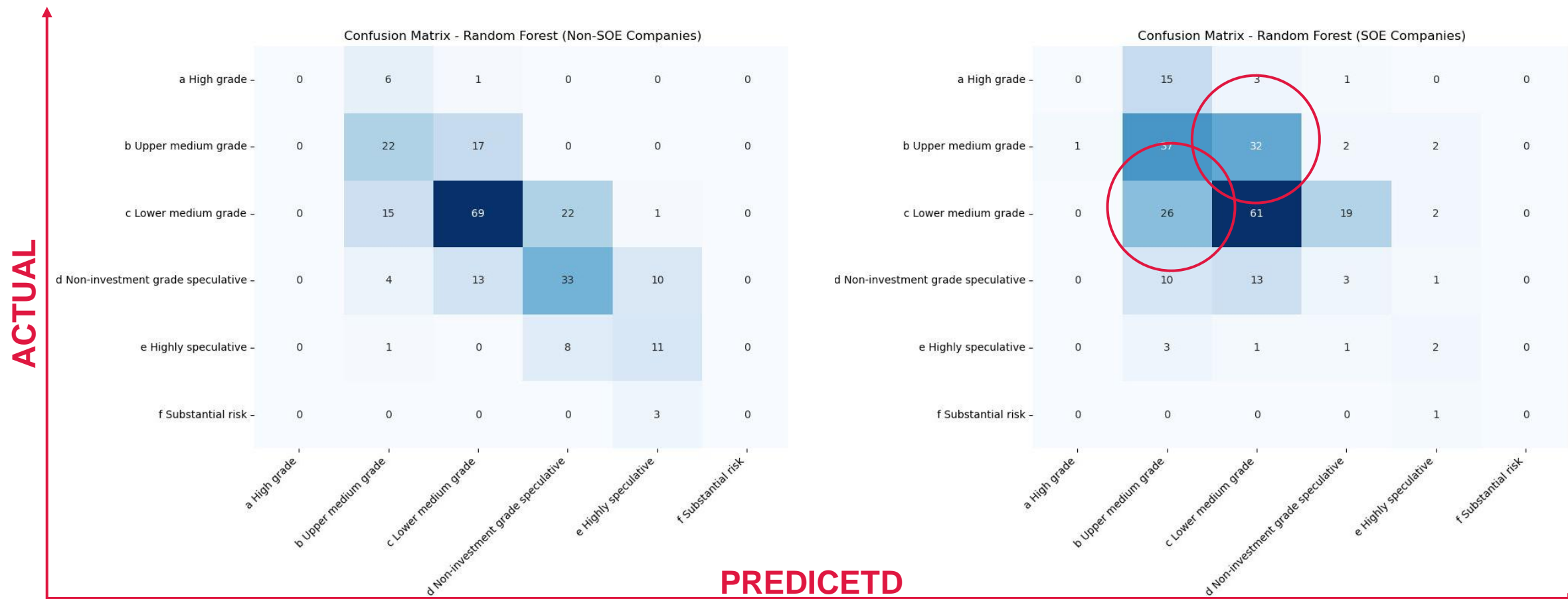
It tells the model:

- “Don’t treat all mistakes equally. If you misclassify a rare class like CCC, that should count as a bigger mistake than misclassifying a common class like A.”

Number of trees: (**n_estimators**) **2,000**

- Each tree is trained on a random subset of the training data (bagging).
- Considers a random subset of features at each decision split.

Confusion Matrix Random Forest: here the mistakes are more balanced.

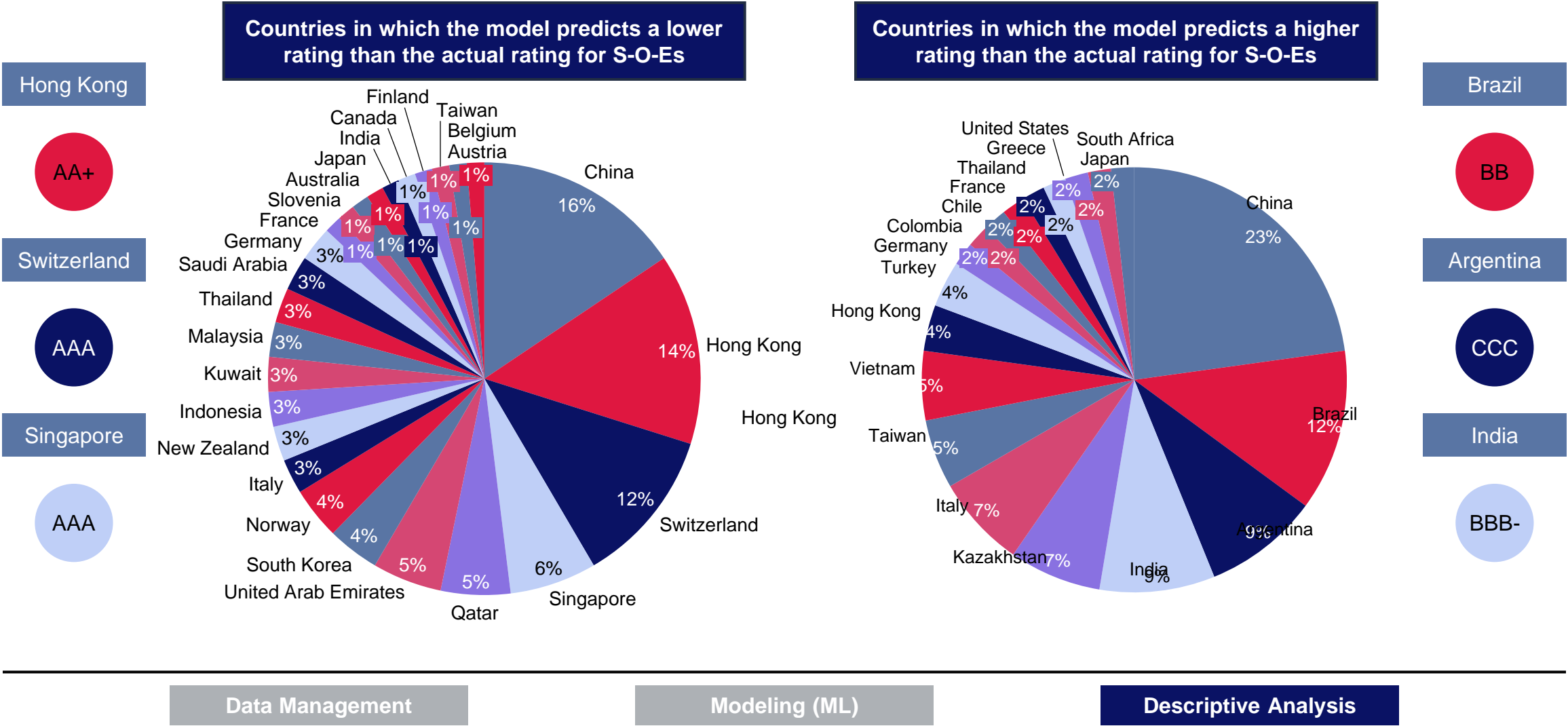


Data Management

Modeling (ML)

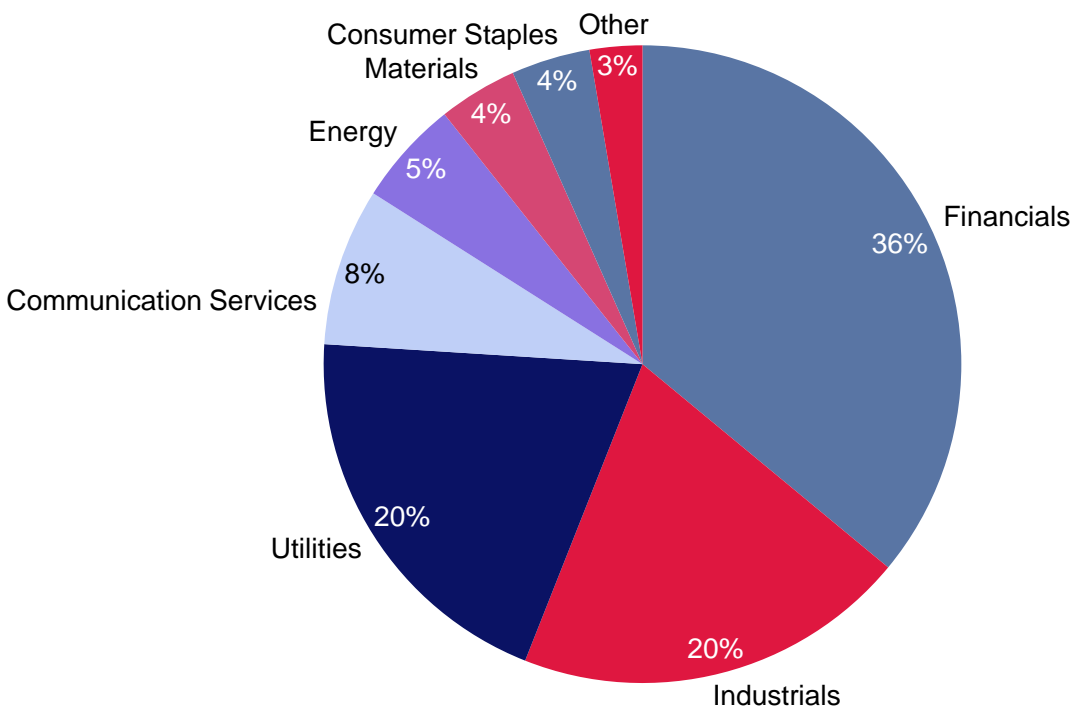
Descriptive Analysis

Model - Random Forest: which countries are uplifting the rating and which countries have a downgrading effect?



Model - Random Forest: which industries are uplifting the rating and which industries have a downgrading effect?

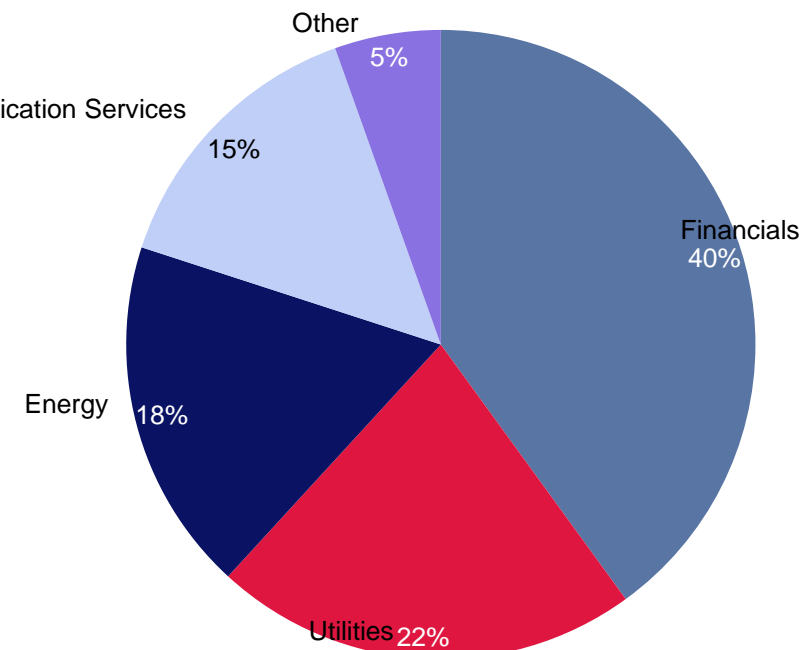
Sectors in which the model predicts a lower rating than the actual rating for S-O-Es



Δ Industrials: 20%
 Δ Utilities: -2%
 Δ Communication Services: -7%
 Δ Energy: -13%

Note: Industrials include transport & logistics, defense and construction.

Sectors in which the model predicts a higher rating than the actual rating for S-O-Es



Data Management

Modeling (ML)

Descriptive Analysis

FLASHBACK:

Are financials less important when determining the rating of State-Owned-Enterprises?

I. Data Retrieval

- i) Bloomberg <EQS> we aim for 3000 companies (already have +- 2500 with only S&P) and 20 financials as well as the text of the ultimate parent, country and industry.
- ii) Cleaning data duplicates between Moody's and S&P, outliers (in terms of financials)
- iii) Text analysis (using GPT) identify S-O-Es extract the sample (say 300 companies and the same number randomly of private companies)

II. Data Visualization

II. Data Visualization

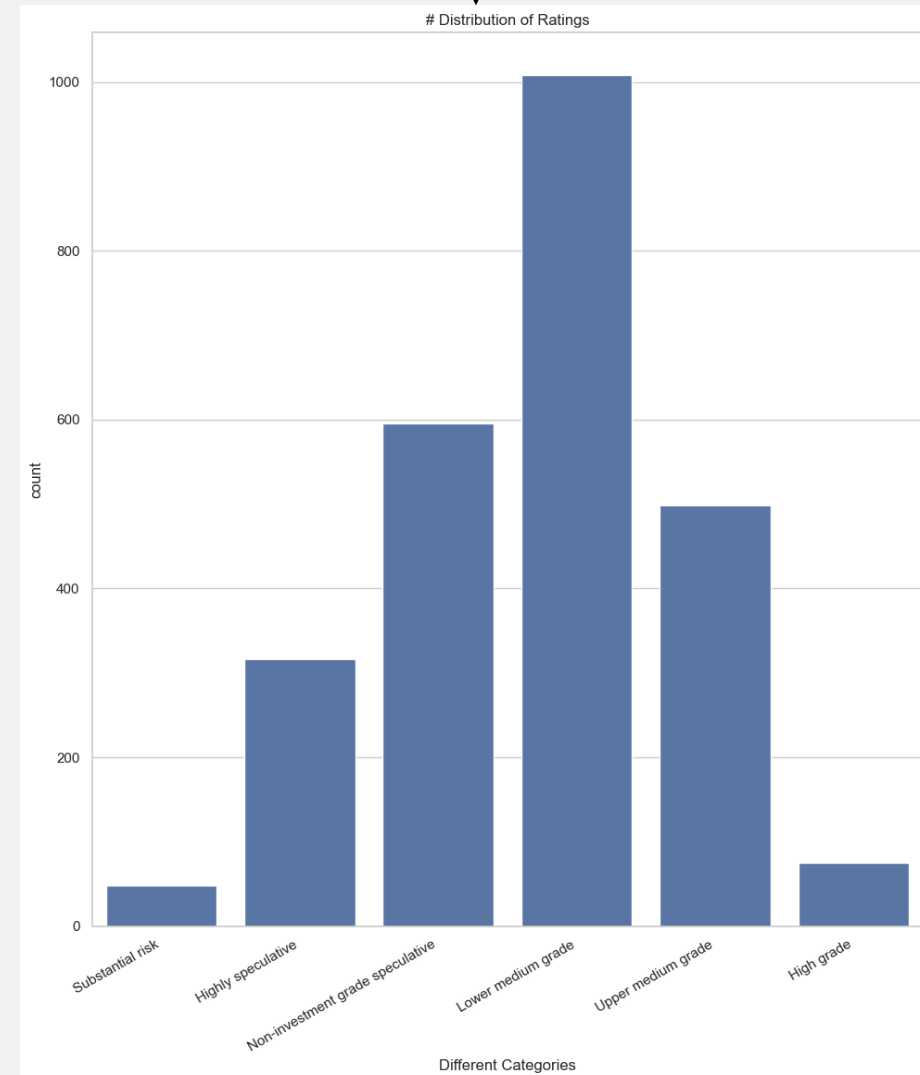
- i) Government related companies' distribution per country
- ii) Government related companies' distribution per company
- iii) Results where the companies have been assigned compared to the real category – for government companies
- iv) Results where the companies have been assigned compared to the real category – for private companies

=> Ideas for the 4 plots

III. Data Modelling

- i) Logistic regression (multinomial (softmax) logistic regression). solver='lbfgs', 'saga' - K-th sample splitting
- ii) Results: confusion matrix
- iii) K-Means Clustering
- iv) Conclusions: Are S-O-E treated similarly as private enterprises? Are they benefitting or not from the state and then **groupby**. (the mistakes in + or -) Country and Industry. Only plotting the errors (+ or -)

Current Data Distribution



Thank you!

Alessandro Fosselard
Antoine Ameye

May 2025

