COLUMBIA ENGINEERING

Commercial Credit Risk Modeling

Author: Nicholas Bindela
Advisor: Miao Wang, Ali Hirsa, Noah Dawang

ask2.ai

Abstract

The analysis of credit risk for loan applicants is important in financial risk management. Being able to confidently predict which loan applicants will fail to repay their loans in full, i.e. default, would reduce much of the present risk associated with lending. Credit risk analysis can be divided into two sectors, commercial and individual. In this project we are interested in analyzing credit risk for commercial loans.

We use real world data, sourced from Moody's Default and Recovery Database (DRD), to build a Machine Learning (ML) model that can predict credit risk for commercial loan applicants. Explainability of the model is analyzed so customers can understand the predictions made.

Project Introduction

Global commercial lending is currently valued at 10.71 Trillion USD and is expected to grow to 21.3 Trillion USD by 2031 [1].

Lenders typically rely on commercial credit ratings to determine whether to extend credit and under what terms. Credit ratings are calculated using various factors that assess a business's creditworthiness.

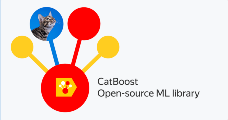
The rise of advanced computing power, innovative algorithms, and the availability of vast amounts of data have fueled the success of complex machine learning (ML) models.

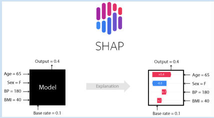
Despite the effectiveness of these "black-box" models, their lack of transparency presents challenges for the various stakeholders involved.

Different stakeholders in the credit risk domain have varying needs for transparency and explanations. Business loan applicants may seek guidance on how to improve their creditworthiness, while lenders need to justify their decisions. Regulators, on the other hand, must ensure that models are fair and function appropriately under extreme conditions.

For this project I trained various tree and gradient boosting machine learning models, to predict whether businesses will default on their loan. Additionally, SHapley Additive exPlanations (SHAP) is used to explain the model's decision.

The CatBoost gradient boosting decision tree model in conjunction with SHAP can be used as a way to confidently predict whether a business loan applicant will default and provide reasons for loan approval or denial.





Data

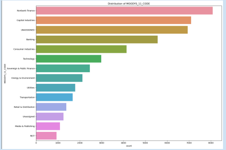
The data used to train the models was retrieved from Moody's Default and Recovery Database (DRD). The DRD contains data for 550,000 debts and 60,000 global sovereigns and corporate entities, including Real Estate, Insurance, and Financial Institutions.

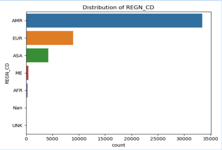
When building the model, the data in the database was aggregated so that each data point contained attributes related to a single business. Our goal is to classify businesses as Default or Non Default based on whether they defaulted on any of their debts.

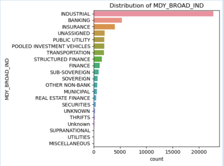
The following aspects for each business were used when building the model:

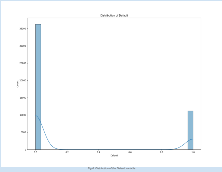
- **Industry Classification**
  - Moody's defined board and specific industry categories.
- **Business Sector Indicator**
  - Moody's categorizes business into one of the general business sectors (banking, industrial, sovereign etc).
  - Moody's also defines sub categories for each sector that are more narrowly defined.
- **North American Industry Classification Code (NAIC)**
  - 3-digit industry code which is a standard used by Federal statistical agencies in classifying business establishments for the purpose of publishing statistical data related to the U.S. business economy.
- **Region Code**
  - Primary global region where the business operates (America, Europe, Asia, etc.).
- **Total and average debt amount held by the business.**
  - All past and present debts were used in the calculation of these values.
- **Coupon rate(s) for debt(s) held by the business**
  - Coupon rate is the annual interest paid by the business for a debt.
  - All past and present debt's rates were included in the model.

Figs 3-6. show the distribution of some of the features included in the model.

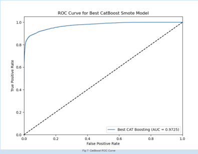








Model Training

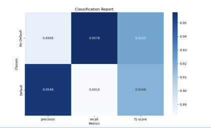


The following methods were used for building models on the data: Logistic Regression, Random Forest, LightGBM, XGBoost, and CatBoost.
Grid search was used to find the optimal parameters for each method.
CatBoost had the highest overall performance with the following parameters:
α (learning rate): 0.15
Iterations : 150
Subsample: 0.5
Depth: 5

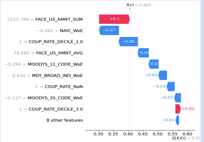
The CatBoost model resulted in an AUC of 0.9725 *[See Fig 7.]*. *Fig 8.* shows the entire classification report.

The **precision** tells us that the model was correct when predicting Non Default 88.88% of the time and 95.48% when predicting Default.

The **recall** tells us that the model correctly identified 95.78% of all Non Defaults and 88.16% of Defaults.



SHAP Input Explainability



SHAP allows us to visualize how features affect the model's classification prediction for a given input.

Fig 9. ranks all features based on the magnitude of affect each has on the model's prediction for a given sample.

The f(x) = 0.405 means that the model predicts there is a 40.5% chance the sample is in the Default class.

Equation 1 shows how the SHAP values for each feature result in the output produced by the model.

$$f(x) = E[f(x)] + \sum_i \phi_i$$

f(x) is the probability the sample is in the class Default as predicted by the model.
E[f(x)] is the average predicted probability of the Default class by the model.
Σϕ is the sum of SHAP values for all features of the given sample.

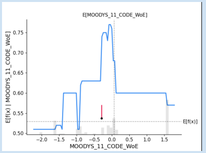
For example, using the sample in *Fig 9.*, the FACE_US_AMT_SUM, which is the sum of the business's debts in USD, of $2,223.784 million influenced the model by 10% towards classifying the sample as a Default.

While the MOODYS_11_CODE of -0.294, which represents the Technology industry *[See Fig. 11]*, influenced the model by 3% towards classifying the sample as a Non Default.

Feature Explainability

SHAP also gives us the ability to visualize how the model's prediction changes based on the value of a specific feature.

Fig 10. shows how the Moody's Industry Classification code impacts the model's prediction.
The y axis represents the probability of Default given the industry denoted on the x axis.



MOODYS_11_CODE	MOODYS_11_CODE_WoE
Unassigned	-2.044445217182110
Nonbank Finance	-1.5932117955392800
Utilities	-1.3786646037966100
Sovereign & Public Finance	-0.9943174503868300
Banking	-0.9017288586786840
Technology	-0.2941006136774580
REIT	-0.13589722880918500
Consumer Industries	-0.0897677257358091
Energy & Environment	-9.4556613647468E-05
Media & Publishing	0.00013830942739981400
Capital Industries	0.0352880368448582O
Retail & Distribution	0.1157657642154570O
Transportation	1.5527175986518800
UNASSIGNED	1.5952746495186700

Fig 11: Moody's 11 encoding mapping

The shape of the graph in *Fig 10.* indicates that the Industry has a large influence on the prediction of the model.
Using *Fig 11.* in conjunction with *Fig 10.* we can see that when given certain industries, notably Technology and Consumer Industries, the model has a much higher probability of predicting a sample will Default.

Conclusion

The CatBoost model trained on the Moody's DRD gives us a way to reliably predict whether an issuer will default based on:

1. **The industry(s)** the business operates in as classified by Moody's and the North American Industry Classification code (NAIC).
2. **The global region** the business primarily operates.
3. The total history of the **business's debts**. This includes the total and average amount of debt held in USD and the individual coupon rates for all of the business's present and past debts.

The model used in conjunction with SHAP allows us to visualize how each feature impacts the model's predictions.

We can also analyze specific samples by ranking features based on how the sample's value for the feature impacted the model's classification.

References

[1] Skyquest (2024). Commercial Lending Market Industry Forecast 2024-2031 (SQMIG40A2010) https://www.skyquestt.com/report/commercial-lending-market#:~:text=The%20Global%20Commercial%20Lending%20Market,period%20(2024%2D2031).
[2] CatBoost - Open-source gradient boosting library. (2024). Catboost. https://catboost.ai/
[3] SHapley Additive exPlanations (SHAP). (2024). Welcome to the SHAP documentation. https://shap.readthedocs.io/en/latest/index.html#welcome-to-the-shap-documentation