QF634 Applied Quantitative Research Methods Project

Singapore Management University

Lee Kong Chian School of Business

Beyond Credit Ratings: A Machine Learning Approach to Real-Time Default Probability Prediction

Authors:
Chia Jun Xian Edmund
Eko Widianto
Anirudh Krishnan
Yohanes Alfredo Phoa

Supervisor:
Prof. Lim Kian Guan

December 16, 2024

Abstract

This study investigates the potential of machine learning (ML) models, particularly tree-based algorithms, to predict implied default probabilities in real time using financial ratios and macroeconomic data. By leveraging market-based Credit Default Swap (CDS) spreads and continuous modeling, the research addresses limitations in traditional credit risk models, such as reliance on static credit ratings and binary classifications. A comprehensive dataset, including CDS prices and financial metrics, was analyzed using LightGBM, XGBoost, and neural networks, with hyperparameter optimization enhancing predictive accuracy. Results demonstrate that tree-based models outperform neural networks, providing stable and precise estimations of default probabilities. The findings highlight the advantages of continuous, market-driven default prediction models in improving risk management, pricing of credit instruments, and regulatory compliance. This study underscores the role of dynamic ML models in evolving credit risk assessment frameworks to adapt to volatile financial environments.

Contents

Chapter 1

Introduction

Default probability is a crucial measure for assessing the likelihood that a borrower will fail to meet their financial obligations. It plays a significant role in risk assessment and management, allowing lenders and investors to evaluate the creditworthiness of borrowers. As a key input in various financial models and instruments, default probability is essential to calculate credit value adjustment (CVA), pricing credit default swaps (CDS), determining capital adequacy requirements, and overall creditworthiness assessment. Consequently, accurate estimation of the probability of default is vital for effective financial decision making and risk management.

However, default probability is not directly observable in the market, presenting significant challenges for its quantification. Although many rely on credit ratings provided by major credit rating agencies (CRA) such as S&P, Moody's, and Fitch, these ratings have notable limitations. CRAs can be subject to herding behavior, where they follow each other's assessments, potentially leading to inflated or delayed ratings. Moreover, there is an inherent conflict of interest, as CRAs are compensated by the issuers of the securities they rate, which may introduce biases into their evaluations. Furthermore, CRAs have historically been slow to respond to market shifts, as seen during the 2008 monetary crisis, when their delayed downgrades of financial institutions were heavily criticized. Furthermore, credit ratings tend to provide generalized assessments and do not account for the specific risk profiles of individual entities, limiting their ability to reflect current market conditions and accurately estimate default probabilities.

In contrast, Credit Default Swaps (CDS) offer an alternative that can provide more time-efficient and market-driven estimates of the probability of default. A CDS is a financial derivative that provides protection against the credit risk of a reference entity, which functions much like an insurance contract. The buyer of a CDS pays a premium to the seller in exchange for protection against a credit event, such as default or failure to pay. The pricing of CDS contracts reflects the collective market assessment of the default probability of the referenced entity, offering a market-implied real-time indication of credit risk.

This study aims to leverage machine learning models to predict the implied default probability of a company using financial ratios and exogenous factors as key input. Using machine learning techniques, we can capture complex relationships between factors and default probability, offering more dynamic and accurate predictions compared to traditional binary models. The goal is to move beyond simple default classification into a continuous, probabilistic prediction of default risk that adapts to market changes.

Chapter 2

Literature Review

Traditional models in credit risk prediction treat default as a binary event — either default (1) or no default (0). These models estimate the probability of belonging to a certain class, such as a credit rating category (e.g., investment grade or junk), from which the default probability is indirectly inferred. Many classical models rely on ratings provided by credit agencies, which categorize entities into predefined classes based on various financial and macroeconomic factors. These ratings are then used to estimate default probabilities. However, this indirect approach often fails to capture the continuous and dynamic nature of default risk, as the class assignment itself is a discrete outcome.

Majority of research in default prediction has focused on binary/multiclass classifications. Studies such as the Credit Research Initiative (CRI) National University of Singapore [1] and research by Moscatelli et al. (2020) [2] focused on such categorization process. While these methods are widely used and offer a practical solution to risk assessment, they often fail to provide a precise, granular view of default risk, as they cannot account for the varying degrees of creditworthiness within each class.

In contrast, the approach proposed in this study predicts implied default probability as a continuous variable, reflecting a more nuanced view of credit risk. Rather than categorizing companies as either defaulted or non-defaulted, continuous models offer a nuanced view that accounts for the varying degrees of creditworthiness over time. This approach allows us to estimate the implied probability of default at any given time, providing a more flexible and granular understanding of credit risk.

Continuous models of implied default probability estimation offer several key advantages over traditional binary models. First, they provide dynamic sensitivity, enabling a more detailed understanding of how implied default probabilities evolve over time. As market conditions change, CDS spreads react to shifts in investor sentiment, macroeconomic factors, and company-specific events, offering a more responsive measure of credit risk, particularly in volatile market environments. Second, continuous models leverage market-based data, with CDS spreads reflecting real-time investor perceptions of default risk. Unlike credit ratings, which are often slow to adjust, CDS spreads provide an up-to-date, market-implied probability of default, making them a timelier and more accurate tool for assessing creditworthiness. Moreover, continuous models enable improved forecasting by capturing the gradual changes in default risk over time, rather than classifying companies simply as "default" or "no default." Finally, continuous models allow for real-time prediction, as they can be continuously updated based on new market data, ensuring that default probability estimates remain current and relevant to investors and financial institutions.

Chapter 3

Implied Default Probability Prediction

3.1 Features Description & Engineering

The data used in this study is sourced from Kaggle [1], which provides the quoted time series of 1–10-year CDS prices from 2015 to 2021, including the period impacted by the COVID-19 pandemic. Since the 5-year CDS is most liquid, the quotes were used to compute the market-implied default probability. The formula used is:

$$\text{Hazard Rate} = \frac{\text{CDS Spread}}{1 - \text{Recovery Rate}} \tag{3.1}$$

$$\text{Probability of Default} = 1 - e^{-\text{Hazard Rate} \cdot \text{time}} \tag{3.2}$$

The recovery rate is assumed to be 40% [3], according to Moody's. Additionally, financial ratios for the respective companies were retrieved from the WHARTON Research Data Services (WRDS) database, which offers comprehensive financial and accounting data.

The financial ratios are scaled using StandardScaler for numerical features, while categorical features like gicdesc and ffi5_desc were transformed using one-hot encoding.

[1] https://www.kaggle.com/datasets/debashish311601/credit-default-swap-cds-prices

Categories	**Ratios**
Profitability Ratios	Price_Sales (ps), Price_Cash_flow (pcf), Net_Profit_Margin (npm), Operating_Profit_Margin_Before_Depreciation (opmbd), Operating_Profit_Margin_After_Depreciation (opmad), Gross_Profit_Margin (gpm), Pre_tax_Profit_Margin (ptpm), Return_on_Assets (roa), Return_on_Capital_Employed (roce), After_tax_Return_on_Average_Common_Equity (aftret_eq), After_tax_Return_on_Invested_Capital (aftret_invcapx), After_tax_Return_on_Total_Stockholders_Equity (aftret_equity)
Solvency Ratios	Long_term_Debt_Invested_Capital (debt_invcap), Total_Debt_Invested_Capital (totdebt_invcap), Capitalization_Ratio (capital_ratio), Cash_Balance_Total_Liabilities (cash_lt), Total_Debt_Total_Assets (debt_at), Short_Term_Debt_Total_Debt (short_debt), Long_term_Debt_Total_Liabilities (lt_debt), Total_Debt_Total_Assets (debt_assets), Total_Debt_Equity (de_ratio)
Efficiency Ratios	Asset_Turnover (at_turn), Sales_Invested_Capital (sale_invcap), Research_and_Development_Sales (rd_sale), Advertising_Expenses_Sales (adv_sale), Labor_Expenses_Sales (staff_sale), Accruals_Average_Assets (accrual)
Market and Company Information	mktcap, price, gicdesc, ffi5_desc, ffi10jj_desc, ffi12_desc, ffi38_desc

Table 3.1: Financial Ratios and Categories

The following macro information are included:

Ticker	Description
^VIX	CBOE Volatility Index
^TNX	10-year U.S. Treasury Yield
^IRX	2-year U.S. Treasury Yield
^TYX	30-year U.S. Treasury Yield
^FVX	5-year U.S. Treasury Yield
TLT	iShares 20+ Year Treasury Bond ETF
LQD	iShares iBoxx $ Investment Grade Corporate Bond ETF
HYG	iShares iBoxx $ High Yield Corporate Bond ETF
MUB	iShares National Muni Bond ETF
TIP	iShares TIPS Bond ETF
EMB	iShares J.P. Morgan USD Emerging Markets Bond ETF
IGOV	iShares International Government Bond ETF

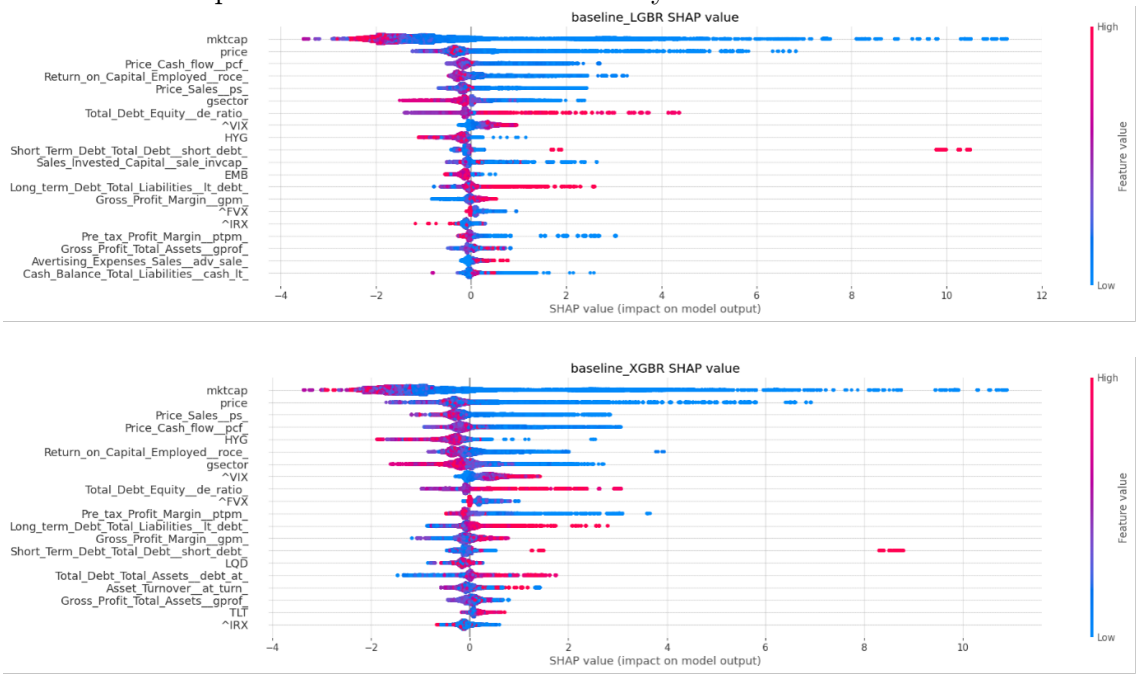
Table 3.2: Financial Ticker Descriptions

The Data was then merged by company and date, dropping rows with NAs to arrive at the final dataset, ready for modelling.

3.2 Modelling Analysis & Results

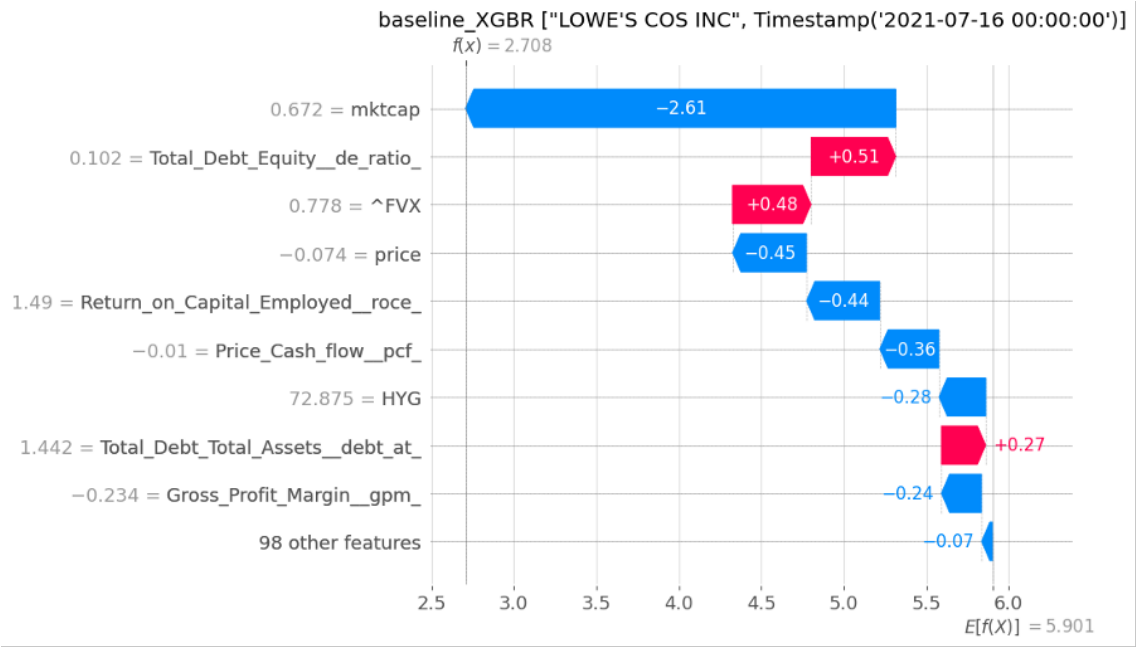
The data were split into training (70%) and testing (30%) sets. Since the dataset is a multi-indexed time series, we avoided shuffling to prevent introducing lookahead bias.

Initial training was performed using baseline LightGBM and XGBoost models, with feature importance and contributions analysed via SHAP values.



The beeswarm plot revealed that market capitalization, price, and price-to-sales were the most important features in predicting implied default probabilities in both models. Companies with lower values for these features were associated with higher default probabilities.

To further investigate feature impact, a waterfall chart was used for a randomly selected datapoint, Lowe's Co. With a market capitalization 0.672 standard deviations above the mean (scaled using StandardScaler), the implied default probability decreased by 2.61%. Conversely, a total debt-to-equity ratio 0.102 standard deviations above the mean increased the implied default probability by 0.51%. These findings align with the theoretical expectation that larger companies, as indicated by higher market capitalization, are financially stronger, while a higher total debt-to-equity ratio increases the likelihood of default.



After training the baseline models, Optuna was used for hyperparameter tuning via Bayesian optimization. In each trial, a set of hyperparameters is suggested, the model is fitted using cross-validation, and the mean squared error is calculated as the loss function. The trial with the lowest mean squared error is selected. The tuning results for the tree-based models are:

Model	**Hyperparameter**	**Value**
LGBR	num_leaves	31
	learning_rate	0.0427
	n_estimators	209
	max_depth	7
	subsample	0.9777
	colsample_bytree	0.6228
XBGR	max_depth	7
	learning_rate	0.0386
	estimator's	274
	subsample	0.8548
	colsample_bytree	0.6149
	min_child_weight	1
	gamma	0.1185

Table 3.3: Hyperparameters for LGBR and XBGR models

A neural network was also trained with the following architecture and optimizer settings:

Layer	Type	Units	Activation
1	Dense	256	ReLU
2	Dropout (10%)	-	-
3	Dense	126	ReLU
4	Dropout (10%)	-	-
5	Dense	1	-

Table 3.4: Neural network architecture.

The model was compiled with the Adam optimizer and using mean squared error as the loss function. In addition, we applied the same optimization framework as the tree-based models, varying the number of hidden layers and units in each layer. Tuning results for the neural-network model:

Layer	Type	Units	Activation
1	Dense	354	ReLU
2	Dropout (10%)	-	-
3	Dense	242	ReLU
4	Dropout (10%)	-	-
5	Dense	82	ReLU
6	Dropout (10%)	-	-
7	Dense	50	ReLU
8	Dropout (10%)	-	-
9	Dense	1	-

Table 3.5: Neural network architecture (hyperparamater tuned).

With the various models, the results are compiled as below:

	Baseline LGBR	Baseline XGBR	Tuned LGBR	Tuned XGBR	Baseline NN	Tuned NN
CV Mean	14.118	14.096	12.695	12.276	27.727	30.851
CV Std	8.802	8.703	10.055	9.589	33.056	42.637

Table 3.6: Comparison of cross-validation results for different models.

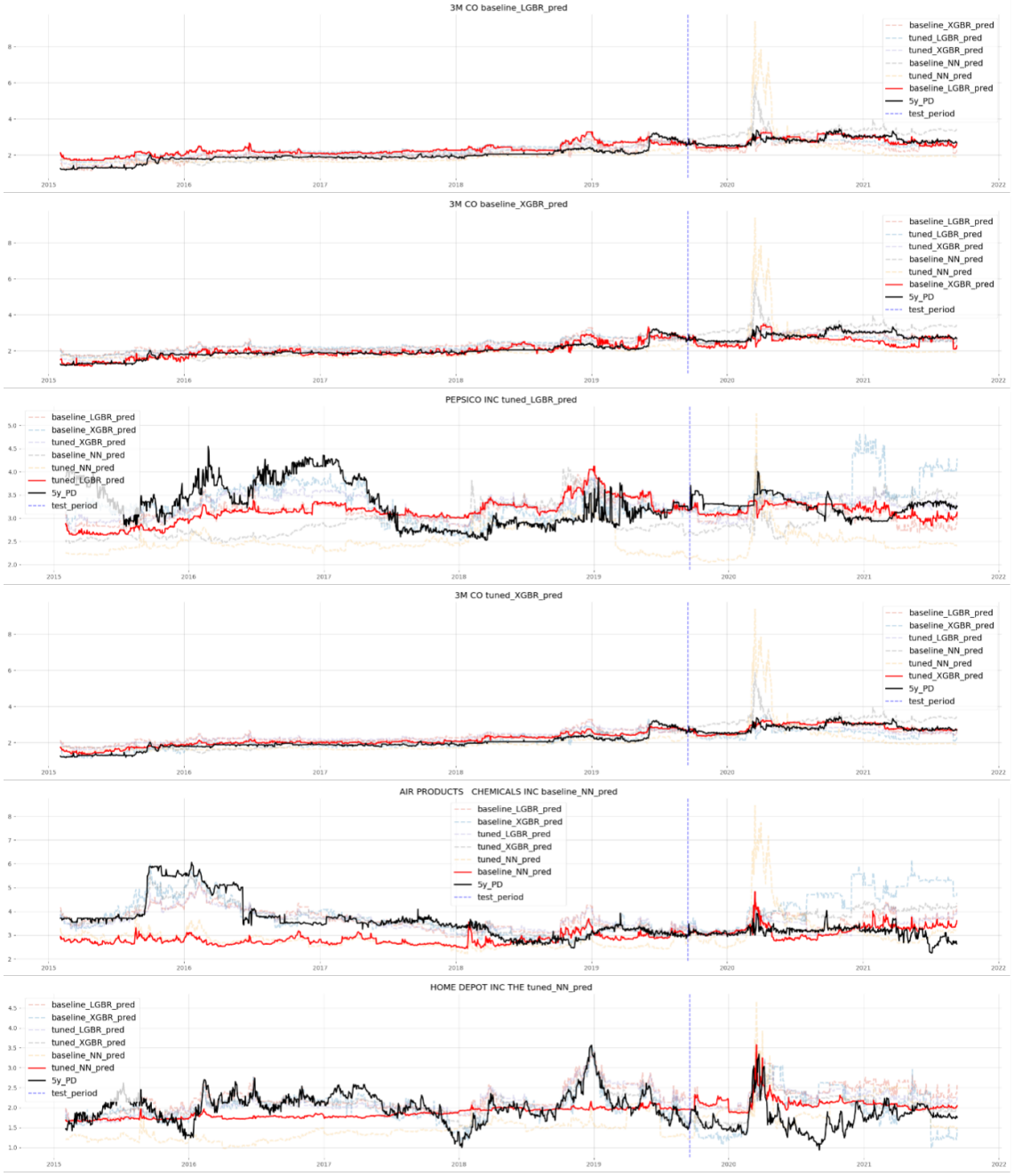
	Baseline LGBR	Baseline XGBR	Tuned LGBR	Tuned XGBR	Baseline NN	Tuned NN
Train	0.59	0.221	0.666	0.282	4.808	4.957
Test	15.408	14.427	13.201	13.307	17.036	20.386

Table 3.7: Train and test results for different models.

Tree-based models consistently outperformed neural networks in both cross-validation and test sets, as indicated by lower mean squared errors and smaller standard deviations during cross-validation. Although hyperparameter tuning improved the performance of tree-based models during cross-validation (lower cross-validation means), it also resulted in higher variability, as seen in the increased cross-validation standard deviations. In contrast, the performance of neural networks, even when tuned, was significantly worse, with much higher cross-validation means and standard deviations.

Overall, tree-based models demonstrated better predictive stability and effectiveness in default prediction, while neural networks showed poorer performance, particularly in terms of cross-validation stability.

It appears that model performance may vary across different companies. Therefore, the best-performing model for each company's testing period should be used to predict its implied default probability moving forward.



Chapter 4

Conclusion

This study highlights the potential of machine learning models, particularly tree-based models, in accurately predicting implied default probabilities. Accurate implied default probability estimation is crucial for effective risk management, enabling financial institutions to make informed lending and investment decisions. It also plays a vital role in pricing financial derivatives, such as Credit Default Swaps (CDS), and calculating Credit Value Adjustment (CVA), which directly impacts capital adequacy requirements and overall portfolio management.

By leveraging financial ratios, macroeconomic factors, and CDS data, this study provides a continuous, market-implied default probability that adapts to market conditions, offering more dynamic and up-to-date insights than traditional binary models. The ability to predict default probability with higher accuracy and responsiveness allows for better decision-making, especially in volatile market environments where timely adjustments are necessary.

In conclusion, the accurate prediction of default probability not only enhances risk assessment but also improves the pricing of credit-related instruments, supports regulatory compliance, and aids in capital allocation decisions. The findings suggest that tree-based models, due to their predictive stability and effectiveness, should be considered as valuable tools in the ongoing effort to enhance credit risk assessment frameworks.

Bibliography

[1] NUS Credit Research Institute. Probability of Default. https://d.nuscri.org/static/pdf/Probability%20of%20Default%20White%20Paper.pdf, 2022. [Online; accessed 29-Nov-2024].

[2] Mirko Moscatelli, Fabio Parlapiano, Simone Narizzano, and Gianluca Viggiano. Corporate default forecasting with machine learning. https://www.sciencedirect.com/science/article/abs/pii/S0957417420303912, 2020.

[3] Moody's Investor Service. Rating methodology: Global financial institutions. https://www.moodys.com/sites/products/defaultresearch/2006600000428092.pdf, 2014. [Online; accessed 29-Nov-2024].