# A Deep Learning Approach to Default Risk Migration Matrices with LLM-encoded Data

Yaning You

*Department of Statistics & Actuarial Science, University of Western Ontario*

**Abstract**

Will finish after proposal itself is fully edited

## 1 Introduction

Interest in capital adequacy assessments have grown exponentially since the signing of the Basel II accord in 2004, shortly after the dot com bubble burst where investment behavior was less than prudent. As time progressed and the world struggled through the Great Financial Crisis of 2008, interest in accurate default forecasts to assess financial capability only grew. As an central component of assessing obligor credit quality, pricing credit related financial instruments such as bonds and credit default swaps, and measuring impact of regulation, default forecast methodologies that are accurate, explainable and robust to the dynamic financial landscape can significantly benefit the operational consistency of financial systems across the world.

### 1.1 Motivation

Corpoate defaults impact the financial system beyond the obligor itself. It represents an immediate financial loss in the company's investors, lenders, and depending on the number of defaults in a time period, the economic loss to the country the company is based in. Understanding the likelihood of a company defaulting helps creditors screen for potential borrowers for more secure returns, investors price instruments at fair value and manage portfolios, and regulators assess the state of the market and decide whether or not to intervene. Thus models of assessing default risk were created by these parties to quantify the risk and track it over time, one of the most popular models being credit ratings.

Credit ratings provide a relatively objective assessment of a firm's ability to repay its existing debt based off of capital structure, historical performance, internal synergies and external influences on a letter scale. This perception of reliability in turn influences lending capacity, investment instrument pricing, corporate strategy, and regulatory oversight of the firm as previously mentioned. Credit ratings gained popularity and became one the most common methods of risk assessment due to two factors. Firstly, it is constructed mainly from firm's financial and economic fundmanetal characteristics, lending the ratings to be highly explicable. More importantly, the ratings are arranged into a letter scale from the highest quality of AAA to companies already in default at D, allowing quick interpretation and decision making, providing a boon for portfolio managers who often face pressure to make high-impact decisions on a tight schedule. Naturally it has been adopted by many in risk modelling and investment management.

The Great Financial Crisis of 2008 (GFC) exposed several weaknesses of the credit rating system as a investment decision making tool. The letter gradings are still too coarse to assess the credit quality of a firm outside of screening purposes. Furthermore, credit ratings are assigned relative to other companies within each rating agency; this not only makes aggregration of firms difficult, but also does not clearly translate to the actual risk of default aside from observing historical rates experienced in each category and performing extrapolations. This means credit ratings are a lagging indicator and are slow to adapt changes in the general state of the economy, particularly changes in the broad environment that may not be directly correlated with the firm, giving the rise in demand for a more granular and objective measure of default risk and the birth of the Probability of Default model, henceforth referred to as PD.

#### 1.1.1 Probability of Default Models in Industry

Unlike credit ratings which assign firms to a finite, often restrictive, set of credit quality indicators, a PD estimates a concrete probability to the firm experiencing financial default. This allows rigorous comparisons between firms across agencies as well as scientific examinations of risk management practices. Due to these factors, PD has quickly gained popularity of usage in credit risk modelling, loan pricing, and portfolio management. There are many ways to model PDs, ranging from a simple logistic regression on observed historical defaults to complex multistep models using firm, market and country-level performance over 10+ years. As of the present, all of the "big three" (S&P, Moody's and Fitch) ratings agencies, as well as many financial data service providers (eg. Bloomberg, Morningstar Refinitiv) are offering PD estimates for firms in its database, a testament to the popularity of PDs in commercial use.

The PD models for each of the three firms is undisclosed; yet there are interesting numances from what information they have revealed. The S&P model uses a dataset of "default flags" (ie. financial statement ratios that can signal default) and derives its PDs with a Maximal Utility approach for optimisation and k-fold Greedy Forward for variable selection bounded by AIC for model parsimony [Chen and Baldassarri, 2015]. Fitch employs

a Monte Carlo simulation model to estimate the probability and frequency of a firm's assets falling below its liabilities, mostly relying on historical credit rating data [Ratings, 2013]. Moody's exhibits the most comprehensive and complex model, using macroeconomic indicators on top of individual firm information, as well as incorporating credit migration matrices to calibrate its model over longer periods of time. The PDs appear to be calculated through a censorship-corrected survival analysis model [Moody's, 2018].

Overall, these models use similar data as the credit ratings and employ relatively straightforward methodologies to model the occurrence of defaults within a credit portfolio. This once again lends to highly explicable PD results that might not fully capture the complexities in the market and broad economy. Naturally, more complex methods of modelling PDs emerged. The proposal will introduce the most straightforward, and once again most popular method, the credit rating migration matrix, in the next subsection, and will provide an overview of 1.2.

### 1.1.2 Credit Rating Migration Matrix

In a portfolio, companies exist in more states than simply default and non-default. The credit rating migration matrix captures a portion of this information by also calculating the probabilities a certain firm improves or degrades in percieved credit quality. Often the rating migrations are mapped to existing credit rating categories for ease of interpretation. The probabilities of migrating into other categories provide an additional level of granularity into corporate default prediction, providing more stable forecasts into future default occurrences, as well as providing stress testing capabilities on existing portfolios. As such the ability to generate accurate, dynamic credit rating migration matrices is crucial to effective pricing and risk management across the financial services industry, motivating the development of credit rating migration matrix forecasting methodologies which will be outlined in 1.2 and 1.3.

## 1.2 Previous Works

The coarseness associated with credit ratings has been addressed as early as the late 60s with Altam's Z score. The function takes multiple solvency ratios, computes a score relative to the firm group using the MDA technique, and firms are ranked using the MDA scores to determine credit risk [Altman, 1968]. Despite being simple to apply and provides more granularity than basic credit ratings, the resulting scores only represent a ranking between the firms and have little intuitive interpretation value otherwise. To counteract this issue, Ohlson introduced a likelihood based forecasting approach, employing logistic regression on similar fundamental financial ratios as input parameter vectors [Ohlson, 1980]. This significantly improves upon MDA based methods by producing a score that directly ties with the probability of default, and a correlation matrix between the input components and probability of default can be generated to analyze for significant parameters. Nevertheless, both methods ultimately only produce a score relative to the data sample, reducing its applicability to broader financial data. Additionally, both methods model probability of defaults at a specific period in time, lending to poor adaptability to changes for forecasts over singular time periods.

Time series-based models arose as a response to this particular issue. Two major approaches of temporal-based analysis arose. Starting with [Shumway, 2001] on a reduced set of market-driven (as opposed to the traditional accounting driven) variables, the first approach involve proportional hazard duration analysis and favored the use of market/index parameters in prediction ([Chava and Jarrow, 2004], [Hillegeist, 2004], [Hull and White, 2000]). Using this work as a theoretical basis, [Duffie et al., 2007] introduced a double stochastic poisson process measuring "time to default" to assist in forecasting, with several advantages. It achieves the main objective of overcoming the compulsory period-by-period forecasting process for the hazard duration analysis models, but also accounts for non-default types of exits such as mergers & acquisitions, and delistings. Unfortunately, the strict high-dimensional data requirements complicate the data formatting process, exponentiate computation time, and introduce parameter instability into the model. [Duan et al., 2012] directly improves upon the double stochastic poisson model, eliminating the previously explicit high-dimensional intermediate state variable estimation through pseudo-likelihood optimisation on CRSP firm and market information. As the function can be decomposed to observe specific parameters for different time periods, it is less computationally intensive and allows for parallel computing. The forward intensity estimate PDs are later used to construct the CRI-PD database used in modelling PD-implied default rates in [Duan and Li, 2021a]. This paper maps PDs from the database to the respective rating grades on the S&P scale based on bounds optimised through data-cloning Sequential Monte Carlo (SMC) techniques [Duan et al., 2020] based on observed historical default occurrences and firm information from CRSP. This introduces a generic technique capable of mapping PD-bsed granular models back to credit ratings, which could encourage the use of PD models in industry for risk analysis and portfolio management. The model was enhanced to use credit rating migration matrices instead of historical observed default rates in [Duan and Li, 2021b]. Using S&P annualized observed migration matrices from the CEREP database filtered out for non-default exits, the implied credit rating migration matrix is calculated for future time periods for each firm of interest. This adjustment solved a shortcoming in the previous methodology where default rates for AAA to AA+ companies had to be arbitrarily extrapolated due to lack of observed defaults, as well as reducing overestimation of companies in the AAA category. However, due to the non-continuous nature of the rating classification, optimisation still relies on SMC, which can degenerate and become computationally intensive [Bourgey et al., 2020].

Introduced in [Li, 1999], the second approach uses gaussian copula functions to study default correlation between companies. Unlike the first method, where the event of default is measured as a discrete probability, the random variable here is "time-until-default". The marginal distributions of firms' survival times are derived from observed market information only, which is

used to determine the joint distribution of survival times for all firms. The model was applied in [Bourgey et al., 2020] to estimate risk on a large credit portfolio and was quickly adapted in industry for calculating Collateralized Debt Obligations (CDOs) due to is ease of implementation and people's familiary with its undelying theory of joint normal distributions [MacKenzie and Spears, 2014]. Yet its simplicity also was the model's biggest pitfall, as its inability to account for tail dependence between parameters resulted in false impressions of CDOs being more diversified then they actually were, leading to the model's subsequent vilification by the public for contributing to the GFC [Zimmer, 2012]. [Allouche et al., 2024] sought to challenge the gaussian coupla model through implementing a temporal-based Dictionary Learning (DL) ML algorithm, constrained by autoregressive regularization. Though the implementation as small-dimension quadratic optimisation problems with linear constraints allow for significantly faster computational implementation compared to the parametric gaussian copula model, and DL is shown to produce more accurate and consistent results compared to the gaussian copula model, the resulting rating migration matrix output consisting of relative signal strengths of rating category upgrades/downgrades is a major setback in interpretability, making the model difficult to use for pricing and regulation purposes in practice. Additionally, the model was trained on incomplete sparse data and was benchmarked against the gaussian copula model on synthetic data, casting doubt on the evident superiority the authors claim.

As computational power improved, there have been studies comparing the performance between ML models and traditional ststistical methods. Studies like [Barboza et al., 2017] and [Sigrist and Leuenberger, 2023] suggested improved prediction and discrimination power for decision tree based models through comparison with traditional statistical techniques. These studies show promise in applying ML methods for default forecasting capabilities; nevertheless there are several major shortcomings of these studies. The biggest flaw among model comparison studies like these come from the statistical benchmarks. Often the machine learning models are compared to rudimentary statistical models such as logistic regression or linear discriminant analysis; it is doubtful the improvement in performance is as significant if the benchmarks were a more complex survival analysis model. In addition, the model results may not be applicable to other markets and economies as their training data is restricted to specific countries in Europe. Deep learning models are also explored in corporate default prediction; [Mai et al., 2019] supports the value in including textual information from 10-K filings in a CNN deep learning framework feeding document-term matrices extracted via Term-Frequency (TF) scanning. [Korangi et al., 2023] leverages transformer-based encoders to create textual data representations appropriate for a multimodal Temporal Convolution Network on mid-cap firms in the S&P index.

### 1.3 Objectives

The approach of this study employs a similar multimodal data processing architecture to that of [Korangi et al., 2023] to forecast default probabilities in the form of credit rating migration matrices, with these following differences. Firstly, the study would be one of the first demonstrations of usaing Large Language Models (LLMs) instead of transformers to encode financial data. LLMs have shown significant potential in human-like levels of textual inference, consistently outperforming traditional pretrained language models such as Bidirectional Encoder Representations from Transformers (BERT) ([Tian et al., 2024]), while also demonstrating decent numerical reasoning abilities with sophisticated prompting [Ahn et al., 2024]. Although there have been methods developed to encode structured data using LLMs ([Perozzi et al., 2024]), there is still a gap in current literature applying these methods using real-world data which this study hopes to begin addressing. Another part where the study is unique is incorporating graph-based Neural Networks to retain complex interactions between inputs that a transformer architecture might be too simple to capture, such as nonlinear boundary conditions in fluid particle simulations ([Horie and Mitsume, 2022], [Zhang et al., 2024]). Finally, the study would like to incorporate macroeconomic indicators to investigate the interaction strength between macroeconomic performance and credit rating migration matrix probabilities. The macroeconomic and country indicators will also be of use when determining model discrimination power and forecast accuracy on training separate credit rating migration matrix forecasting models on each major index of interest (S&P 500, STOXX 600, and TSX Composite) versus training an overarching model on all three indices together.

## 2 Methodology

The paper would break down the methodology for both the LLM encoding and temporal deep learning models, explaining the underlying theory behind each method. For the LLM multichannel data encoding, any prompt templates and prompt engineering techniques used to fine-tune the model will be disclosed, along with justification with how the application of these technqiues assist in improved encoding outputs. For the temporal graph-based neural network implementation, the paper will be structured to first provide a mathematical explanation of the central reasoning architecture, followed by a justification of the advantages of including graph knowledge structures and why this particular model was selected. An application of the methodology will follow in a section discussing experimental design. This section will begin by introducing the datasets used throughout the study and describe the collection procedure, as well as any challenges within the dataset that required a workaround during model training. This proceeds into describing the training procedure, including the computing resources used, data preprocessing steps, the implementation of workarounds (if there were any), and the metrics used to evaluate model output. Finally, the experiment results in both computation effeciency and pre-

dictive power are revealed. This deconstruction will be repeated for standalone model training and comparison with benchmark data. The paper will close by concluding the results of the experiment and the usefulness of the study, addressing all of the hypotheses in section 1.3, conceding on any shortcomings on the study, and propose areas of future investigation.

## 2.1 Data Collection

Currently most of the data gathered for this study comes from S&P Capital IQ. As S&P Capital IQ is a commercial platform and access is not universal, this project will be one of the few studies utilizing this dataset and will rely on S&P data for accuracy benchmarking. Data to train and evaluate the model will be collected in two steps:

1. Construct credit risk data profile for constituents of S&P 500, STOXX 600, and TSX Composite Indices from Capital IQ. The composition of each credit risk data profile is as follows:

   (a) Fundemantal data: this is information regarding individual firms. Data will be collected in these four areas: - Firm volatility (This would involve calculating the % change in the 30 day moving average from previous period)

   - Equity and Fixed income prices (30 day moving averages)

   - Financial information (financial statement items and ratios often used to assess borrowing capability)

   - Total assets

   - Loss provisions

   - Effeciency: asset turnover, cash flow cycle

   - Profitability: Return on equity, return on assets

   - Leverage: Debt to Equity, EBIT to interest expense

   - Liquidity: current ratio, tier 1 capital ratio

   (b) Market data: this is information on the underlying index (so the S&P 500, STOXX 600 and TSX Composite). Data will be collected in these areas:

   - Market volatility (This would be in the form of analyzing the historical performance of the COBOE Volatility Index (VIX) and Euro STOXX 50 Volatility (VSTOXX))

   - Index prices

   (c) Macroeconomic data: this is information on the country the index is derived from. Data would be collected in these areas:

   - real GDP change

   - Volatility (in the form of interest rates)

   - Macroeconomic trends (Considering major news events throughout the years)

2. Gather benchmarks to compare the deep learning model output with input. The model would be benchmarked on both accuracy and performance. For accuracy, one source of benchmarks would be the credit migration matrix results presented by the PDiR paper (insert citation here). The model can be trained to also return PDs of individual index constituents, which can be compared with the PDs generated by S&P Capital IQ. Currently the possibility of benchmarking the model performance with S&P generated credit risk migration matrices is under examination; if such data exists and is accessible this will provide another layer of accuracy validation. Comparison of performance will be discussed in 2

## 2.2

1. Encode the data profiles using LLM into tensors to feed into deep learning model. Current considerations of the transforming LLM models include a Chronos-T5 embedding model to capture temporal relationships, or a General LLM (10B+ parameters) to exploit a more comprehensive knowledge base and can perhaps inference more relevant relationships among the data. Overall, the goal of this step is into transform the firm, equity/debt market and macroeconomic data into a format that can be understood by a deep learning model.

2. Generate predicted credit migration matrices using deep learning model

   - Graph-NN

   - SineNet

3. Compare model results to results produced by existing work

   - If permissible, test the predictive power of training one vs. multiple models

# References


Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

Michael Allouche, Emmanuel Gobet, Clara Lage, and Edwin Mangin. Structured dictionary learning of rating migration matrices for credit risk modeling. *Computational Statistics*, 39:3431–3456, 01 2024. doi: 10.1007/s00180-023-01449-y. URL https://link.springer.com/article/10.1007/s00180-023-01449-y#citeas.

Edward I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4):589–609, 1968. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/2978933.

Flavio Barboza, Herbert Kimura, and Edward Altman. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 83:405–417, 2017. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.04.006. URL https://www.sciencedirect.com/science/article/pii/S0957417417302415.

Florian Bourgey, Emmanuel Gobet, and Clément Rey. Metamodel of a large credit risk portfolio in the gaussian copula model. *SIAM Journal on Financial Mathematics*, 11(4):1098–1136, 2020. doi: 10.1137/19M1292084. URL https://doi.org/10.1137/19M1292084.

Sudheer Chava and Robert A. Jarrow. Bankruptcy Prediction with Industry Effects*. *Review of Finance*, 8(4):537–569, 01 2004. ISSN 1572-3097. doi: 10.1093/rof/8.4.537. URL https://doi.org/10.1093/rof/8.4.537.

A. Chen and G. Baldassarri. Pd model fundamentals: Banks - a pioneer model for assessing bank creditworthiness, 2015. URL https://www.spglobal.com/marketintelligence/en/documents/pd-model-fundamentals-banks-a-pioneer-model-for-assessing-bank-creditworthiness.pdf.

Jin-Chuan Duan and Shuping Li. *PD-Implied Ratings via Referencing a Credit Rating/Scoring Pool's Default Experience*, pages 105–115. Springer International Publishing, Cham, 2021a. ISBN 978-3-030-49728-6. doi: 10.1007/978-3-030-49728-6_6. URL https://doi.org/10.1007/978-3-030-49728-6_6.

Jin-Chuan Duan and Shuping Li. Enhanced pd-implied ratings by targeting the credit rating migration matrix. *The Journal of Finance and Data Science*, 7:115–125, 2021b. ISSN 2405-9188. doi: https://doi.org/10.1016/j.jfds.2021.05.001. URL https://www.sciencedirect.com/science/article/pii/S2405918821000052.

Jin-Chuan Duan, Jie Sun, and Tao Wang. Multiperiod corporate default prediction—a forward intensity approach. *Journal of Econometrics*, 170(1):191–209, 2012. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2012.05.002. URL https://www.sciencedirect.com/science/article/pii/S0304407612001145.

Jin-Chuan Duan, Andras Fulop, and Yu-Wei Hsieh. Data-cloning smc2: A global optimizer for maximum likelihood estimation of latent variable models. *Computational Statistics & Data Analysis*, 143:106841, 2020. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2019.106841. URL https://www.sciencedirect.com/science/article/pii/S0167947319301963.

Darrell Duffie, Leandro Saita, and Ke Wang. Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3):635–665, 2007. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2005.10.011. URL https://www.sciencedirect.com/science/article/pii/S0304405X06002029.

Stephen A. et al. Hillegeist. Assessing the probability of bankruptcy. *Review of Accounting Studies*, 9(1), 3 2004. ISSN 1573-7136. doi: 10.1023/B:RAST.0000013627.90884.b7. URL https://doi.org/10.1023/B:RAST.0000013627.90884.b7.

Masanobu Horie and Naoto Mitsume. Physics-embedded neural networks: Graph neural pde solvers with mixed boundary conditions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23218–23229. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/93476ae409ae3246e22a9d4b931f84ed-Paper-Conference.pdf.

John Hull and Alan White. Valuing Credit Default Swaps II: Modelling Default Correlations. Technical report, Joseph L. Rotman School of Management, University of Toronto, Toronto, 2000.

Kamesh Korangi, Christophe Mues, and Cristián Bravo. A transformer-based model for default prediction in mid-cap corporate markets. *European Journal of Operational Research*, 308(1):306–320, 2023. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2022.10.032. URL https://www.sciencedirect.com/science/article/pii/S0377221722008207.

David X Li. On default correlation: A copula function approach. *Available at SSRN 187289*, 1999. doi: https://dx.doi.org/10.2139/ssrn.187289. URL https://ssrn.com/abstract=187289.

Donald MacKenzie and Taylor Spears. 'the formula that killed wall street': The gaussian copula and modelling practices in investment banking. *Social Studies of Science*, 44(3):393–417, 2014. ISSN 03063127. URL http://www.jstor.org/stable/43284238.

Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 274(2):743–758, 2019. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2018.10.024. URL https://www.sciencedirect.com/science/article/pii/S0377221718308774.

Moody's. Features of a lifetime pd model: Evidence from public, private, and rated firms, 2018. URL https://www.moodys.com/web/en/us/insights/credit-risk/features-of-a-lifetime-pd-model.html.

James A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18 (1):109–131, 1980. ISSN 00218456, 1475679X. URL http://www.jstor.org/stable/2490395.

Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for llms, 2024.

Fitch Ratings. Fitch portfolio credit model, 2013. URL https://www.fitchratings.com/fitch-portfolio-credit-model.

Tyler Shumway. Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, 74(1):101, 2001. ISSN 00219398. URL https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4043236&site=ehost-live.

Fabio Sigrist and Nicola Leuenberger. Machine learning for corporate default risk: Multi-period prediction, frailty correlation, loan portfolios, and tail probabilities. *European Journal of Operational Research*, 305(3):1390–1406, 2023. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2022.06.035. URL https://www.sciencedirect.com/science/article/pii/S0377221722005100.

Jiahao Tian, Jinman Zhao, Zhenkai Wang, and Zhicheng Ding. Mmrec: Llm based multi-modal recommender system, 08 2024.

Xuan Zhang, Jacob Helwig, Yuchao Lin, Yaochen Xie, Cong Fu, Stephan Wojtowytsch, and Shuiwang Ji. Sinenet: Learning temporal dynamics in time-dependent partial differential equations, 03 2024. URL https://arxiv.org/abs/2403.19507.

David M. Zimmer. The role of copulas in the housing crisis. *The Review of Economics and Statistics*, 94(2):607–620, 2012. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/23262091.