Beyond Linear Models: Explainable AI for Merger & Acquisition Outcome Prediction

Master's thesis for obtaining the academic degree

Master of Science

in the study programme Transition, Innovation and Sustainability Environments (TISE)

submitted by

Alexander Ryusandi Pratama

Department for Knowledge and Communication Management

at University for Continuing Education Krems

Advisor: Prof. dr hab. Adam Zaremba

Advisor: Dr. Ruihai Dong

Vienna, 5 June 2025

STATUTORY DECLARATION

I, Alexander Ryusandi Pratama, hereby declare,

1. that I have written my Master's thesis myself, have not used other sources than the ones stated and moreover have not used any illegal tools or unfair means,

2. that I have not used my Master's thesis or parts thereof as an exam paper in my domestic or any foreign country in any form to this date,

3. that, in case my Master's thesis concerns my employer or any other external cooperation partner, I have fully informed them about title, form and content of the Master's thesis and have his/her permission to include the data and information in my written work.

Ryu

Alexander Ryusandi Pratama

Vienna, 5 June 2025

ACKNOWLEDGEMENTS / DEDICATIONS

This thesis, much like the journey across continents it metaphorically mirrors, is not a solitary endeavor but a mix of intellectual currents, personal encounters, and unwavering support that have collectively shaped me. My "searching for love," first penned on a hopeful note, has indeed evolved, from an intellectual journey into a profound realization that the most meaningful discoveries often lie in the process itself: in the rigorous pursuit of knowledge, the embrace of diverse perspectives, and the quiet resilience forged in moments of challenge.

My deepest gratitude extends to my academic mentors, who served as pivotal navigators in this intellectual journey. To Prof. dr hab. Adam Zaremba from *Uniwersytet Ekonomiczny w Poznaniu*, thank you for your unwavering insistence on probabilistic rigor and the elegant precision of modeling; your guidance transformed ideas into defensible arguments. To Dr. Ruihai Dong from University College Dublin, your profound insights into the strategic dimensions of machine learning provided an essential anchor, consistently prompting deeper reflection and a more nuanced understanding. To Koby Shemer from AlphaBeta (I.S.I) Investment Indices Ltd., I am indebted for your focus on practical applicability, for pushing my theoretical explorations into the empirical crucible where abstract concepts must prove their mettle through robust experimentation. Your collective mentorship was the critical scaffolding upon which this work was built.

The model of the Erasmus Mundus Joint Master in Transition, Innovation, and Sustainability Environments (TISE) that is one of a kind in this world, has provided an unparalleled experience academically and holistically. I extend my sincere appreciation to the faculty and staff across the consortium: at the *Universidade NOVA de Lisboa* (NOVA University Lisbon), where my journey into digital cultures began; at University College Dublin, which immersed me in the complexities of technologies; at *Uniwersytet Ekonomiczny w Poznaniu* (Poznań University of Economics and Business), for the crucial study of transition economies; and finally, at the *Universität für Weiterbildung Krems* (University for Continuing Education Krems), where the threads of complexity science wove this interdisciplinary threads together. Mag. Kay Mühlmann and Dr. Liliya Satalkina's collective commitment to fostering a borderless academic community, marked by prompt feedback, thoughtful mentorship, and logistical agility, created a fertile ground for intellectual growth. The financial support from the European Union Erasmus Programme was the very bedrock that made this transformative educational experience possible. My gratitude extends to AlphaBeta (I.S.I) Investment Indices Ltd., for the invaluable internship opportunity. Access to real-world M&A data and the challenge of developing forecasting tools within a professional quantitative finance environment provided a bridge between theory and practice, significantly enriching the empirical foundations of this thesis.

This journey was also profoundly personal. The initial transition, carrying a freshly minted Sociology degree from *Universitas Indonesia* and a pouch of Indonesian spices into the sophisticated world of complexity science, was a formidable leap. It was a path punctuated by moments of self-doubt, such as the visa and resident permit challenges that became an unexpected lesson in resilience. These experiences: quiet reflection in a windowless room amid Lisbon's winds, daily journeys through the traffic and forests in the rain of Dublin and Bray, sleepless nights in the hot and humid air of Porto, enduring the bitter cold of a Soviet-era flat in Poznań, and walking along the flowers in the calm Danube River of Krems an der Donau and Vienna, were pivotal in shaping both my academic resilience and a deeper self-awareness.

Finally, to my family, friends, and the Indonesian Student Association in Portugal, Ireland, Poland, and Austria: your unwavering belief, your patient encouragement across three time zones, the humor that lightened difficult days, and the countless small kindnesses formed an invisible net of support that sustained me through every draft and every doubt. This achievement is as much a testament to your love and sacrifice as it is to any effort of my own.

This work, then, is a culmination, a marker on a path of continuous learning. The search continues, not for a final destination, but for ever-deeper understanding, for connections that illuminate, and for the quiet strength found in embracing the beautiful complexity of our shared human and intellectual experience.

This work acknowledges the Three Laws of Robotics:

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

— Asimov, I. (1950). I, Robot. Gnome Press. pg 9-134

And importantly, the Zeroth Law:

A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

— Asimov, I. (1985). Robots and Empire. Nightfall, Inc. pg 84-91

Abstract

This thesis critically evaluates the efficacy of the Tabular Prior-informed Bayesian Neural Network (TabPFN), for predicting Merger & Acquisition (M&A) outcomes. Addressing the limitations of conventional econometric models and the opacity of "black-box" AI in regulated financial domains, particularly within the Transition, Innovation, and Sustainability Environments (TISE) framework, this research compares TabPFN against benchmark tuned models: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). The analysis utilized a dataset of 4,400 historical U.S. M&A transactions with announcement dates ranging from January 12, 2000, to July 11, 2024, and corresponding exit dates (deal completion or termination) extending from February 14, 2000, to December 26, 2024, characterized by significant class imbalance.

The primary aim was to assess differences in predictive discrimination (AUC-ROC), probability calibration (ECE), and explanatory insights. Empirical findings revealed TabPFN achieved the highest discriminative performance (AUC-ROC: 0.9168), outperforming LR (0.7171) and other baselines. TabPFN also demonstrated strong probability calibration (ECE: 0.0676), significantly better than LR (ECE: 0.1372), although RF exhibited the best ECE (0.0299). SHAP analysis of TabPFN identified 'realized_return' as a key predictor for the tested instances and elucidated crucial non-linear feature effects not apparent from linear models; the initial feature pruning, guided by Random Forest, highlighted the general importance of factors including 'attitude_Friendly' and 'completion_time' for the prediction task.

This study concludes that TabPFN's superior discriminative performance and strong calibration, synergistically combined with XAI techniques, provide a more valuable and actionable tool for M&A strategic decision-making. The enhanced probabilistic reliability and transparent, granular feature-level explanations offer a robust foundation for risk assessment within the TISE framework, aligning with demands for trustworthy AI in high-stakes financial applications.

Keywords:

Mergers and Acquisitions (M&A) Outcome Prediction; Explainable Artificial Intelligence (XAI); Tabular Prior-informed Bayesian Neural Network (TabPFN); Probability Calibration; SHapley Additive exPlanations (SHAP); Transition, Innovation, and Sustainability Environments (TISE) Framework.

CONTENTS

Introduction......1

1.1 Transition-Finance Context......1

1.2 Empirical and Theoretical Problem Gap......2

1.3 Research Aim and Research Questions......3

1.4 Guiding Conceptual Lenses......4

Literature Review......6

2.1 M&A Success Constructs and Measurement Heterogeneity......6

2.2 Limitations of Econometric Baselines......8

2.3 Machine Learning in Corporate-Event Prediction......10

2.4 Transformers for Tabular Inference......12

2.5 Explainable AI in Regulated Finance......15

2.6 Integrative Gap Statement and Resulting Hypotheses......17

Methods......21

3.1 Research Paradigm & Design......21

3.2 Data Source, Sampling, Inclusion/Exclusion, Ethical Compliance......22

3.3 Feature Engineering Rationale......25

3.4 Pre-Processing Pipeline......26

3.5 Model Specifications......26

3.6. Architectural and Inferential Framework......27

3.7 Validation Strategy and Performance Metrics......28

3.8 Explainability Protocol......29

3.9 Robustness & Reproducibility Safeguards......30

Results......32

4.1 Sample Characteristics......32

4.2 Model Discrimination & Calibration......33

4.3 Robustness Checks......35

4.4 Explainability Outputs......35

Discussion......37

5.1 Principal findings mapped to RQs/Hypotheses......37

5.2 Theoretical implications......38

5.3 Comparison with Extant Literature......39

5.4 Methodological Reflections......41

5.5 Practical Implications........ 43
5.6 Strengths & Limitations........ 45
5.7 Future Research Agenda........ 47
Conclusion........ 49
References........ 51
Glossary........ 67
Index/ANNEXES........ 71
A. Ethical Considerations........ 71
B. Categories and NAICS of the Dataset........ 74
C. Deconstructing TabPFN for M&A........ 78
D. Full Script........ 79
E. Results........ 87

LIST OF TABLES

Number/ Alphabet		*Page*
4.1	Comparative Performance Metrics of Primary Evaluated Models on Hold-Out Test Set	34
B	Categories and NAICS of the Dataset	74
C	Deconstructing TabPFN M&A	77

LIST OF FIGURES

Number		*Page*
2.6	Literature Reviews Connections	20
3.2	Categorical Feature Frequencies Example	25
3.9	Methodology	31

ABBREVIATIONS

AI: Artificial Intelligence

ALE: Accumulated Local Effects

AP: Average Precision (Precision-Recall)

AUC-ROC: Area Under the Receiver Operating Characteristic Curve

BaFin: Federal Financial Supervisory Authority (Germany)

CAS: Complex Adaptive Systems

CSV: Comma-Separated Values

CV: Cross-Validation

DOI: Digital Object Identifier

ECE: Expected Calibration Error

ESMA: European Securities and Markets Authority

EU AI Act: European Union Artificial Intelligence Act

FAIR: Findable, Accessible, Interoperable, Reusable

FinTech: Financial Technology

GBM: Gradient Boosting Machine

i.i.d.: independent and identically distributed

LIME: Local Interpretable Model-agnostic Explanations

LR: Logistic Regression

M&A: Merger and Acquisition

MAR: Missing At Random

ML: Machine Learning

MLP: Multilayer Perceptron (Machine Learning Model)

MLP: Multi-Level Perspective (System Theory)

NLP: Natural Language Processing

PFN: Prior-Data Fitted Network

RBV: Resource-Based View

RNG: Random Number Generator

ROA: Return on Assets

SCM: Structural Causal Model

SHAP: SHapley Additive exPlanations

SMOTE: Synthetic Minority Over-sampling Technique

TabPFN: Tabular Prior-informed Bayesian Neural Network

TISE: Transition, Innovation and Sustainability Environments

XAI: Explainable Artificial Intelligence

BEGINNING OF THE TEXT

INTRODUCTION

1.1 Transition-Finance Context

The global economy is undergoing significant transformations due to technological advancements, market instability, and shifting geopolitical dynamics, all of which create complex challenges for industries. These disruptions require companies to reassess strategies and allocate resources effectively to remain competitive. One essential mechanism used in this process is Mergers and Acquisitions (M&A), which involve companies combining through a merger to form a new entity or an acquisition where one firm takes over another. M&A transactions play a crucial role in reshaping markets, consolidating assets, and driving innovation or sustainability efforts. By enabling firms to achieve operational synergies, expand capabilities, or reposition strategically, M&A fosters economic adaptability. However, these transactions are highly complex and prone to failure due to misaligned expectations, integration difficulties, or inaccurate valuation assessments. As a result, precise forecasting and informed decision-making are necessary to minimize risks and improve outcomes in an evolving financial landscape (Gomes et al., 2013); (King et al., 2004).

Within the framework of Transition, Innovation, and Sustainability Environments (TISE), managing the financial aspects of corporate transformations is essential. Transition finance plays a key role in supporting economic activities through financial tools, strategies, and insights that promote resilience and sustainability (Muehlmann et al., 2022; Satalkina et al., 2021, 2022). Efficient capital allocation is particularly challenging in periods of uncertainty and complex, non-linear market dynamics. Traditional M&A forecasting, often reliant on linear assumptions and stable economic indicators, is ill-equipped for the non-linear dynamics and structural uncertainties inherent in transition-finance. Recent research confirms that financial analysts' forecast accuracy sharply declines post-merger due to increased complexity, changes in capital intensity, and information asymmetries (King et al., 2004).

Addressing these limitations requires innovative financial strategies capable of adapting to evolving economic conditions and improving decision-making in corporate transformations. AI-driven approaches, such as transformer-based models and hybrid machine learning systems, significantly outperform traditional methods in predictive performance. For example, transformer architectures like TabPFN have demonstrated advanced capabilities in capturing high-dimensional, nonlinear patterns in tabular M&A data, offering better interpretability and generalization across diverse scenarios (McCarter, C., 2022); (Xu et al., 2024). These methods are especially beneficial in dealing with imbalanced datasets and high-dimensional variables by integrating advanced feature selection mechanisms and probabilistic modeling (McCarter, C., 2022).

This research directly addresses the gap in M&A outcome forecasting by examining the application of innovative, explainable Artificial Intelligence (xAI), specifically the transformer-based Tabular Prior-informed Bayesian Neural Network (TabPFN) model. By significantly improving predictive accuracy and offering deeper, non-linear insights into deal success factors compared to conventional approaches, this study contributes innovative AI-driven tools for corporate finance. Studies also demonstrate that such models are effective not only in forecasting acquisition likelihood but also in estimating post-merger synergy potential, thus aiding capital allocation aligned with long-term value creation (Xu et al., 2024). These innovations enable more precise capital allocation by reducing misallocation in ventures and helping identify M&A transactions that genuinely foster sustainable innovation rather than short-term market consolidation. Ultimately, this approach enhances strategic decision-making, supporting more resilient and sustainable financial strategies amid ongoing economic and industrial shifts (McCarter, C., 2022).

1.2 Empirical and Theoretical Problem Gap

Despite the crucial role Mergers and Acquisitions (M&A) play in economic transitions, they suffer from a persistently high failure rate. Empirical studies reveal that analyst forecast accuracy significantly deteriorates after M&A transactions due to increased organizational complexity, changes in capital intensity, and amplified information asymmetries (Bi & Zhang, 2021). Industry-wide and academic assessments estimate failure rates between 50% and 90%, depending on definitions of failure ranging from unmet shareholder value to poor integration and unrealized synergies (Craninckx & Huyghebaert, 2010); (Rozen-Bakher, 2018); (Mcgaughan, J. A., 2021). Multiple sources confirm that synergy realization is especially elusive in large or cross-border transactions due to integration complexity and misaligned expectations (Larsson & Finkelstein, 1999); (Rozen-Bakher, 2018).

Critically, many M&A deals not only fail to create value but actively destroy it. Post-merger integration problems especially related to human capital, cultural friction, unrealistic synergy assumptions, and inflated acquisition premiums consistently rank among the top reasons for failure (Mirc et al., 2022); (Garzella & Fiorentino, 2017). Even strategically sound deals fail when leadership overestimates integration capabilities or fails to anticipate executional complexity (Ridderheim & Pålsson, 2012).

These failures are often exacerbated by the volatile, non-stationary environments in which M&As occur. Market shocks, regulatory shifts, and evolving technological landscapes weaken the reliability of historical performance data. Compounding this are cognitive biases such as managerial overconfidence and confirmation bias that skew executive judgment, resulting in flawed decision-making and misaligned resource allocation (Urlichs et al., 2014). Moreover, traditional financial models lack the capability to predict outcomes in the highly nonlinear, multivariate environment of M&A performance. These models often underperform due to their inability to integrate soft variables such as cultural fit, employee sentiment, or leadership agility (Fiorentino & Garzella, 2013).

To address these methodological limitations, this thesis investigates the application of explainable Artificial Intelligence (XAI), particularly the transformer-based Tabular Prior-informed Bayesian Neural Network (TabPFN), as an advanced predictive model for M&A outcomes. TabPFN is recognized for its strong performance on small-sample, high-dimensional tabular data, offering a learned Bayesian prior that supports robust uncertainty modeling and non-linear inference without extensive hyperparameter tuning (Hollmann et al., 2023; Müller et al., 2022). This study will compare its efficacy against a suite of tuned baseline models including Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP), focusing on features such as deal characteristics (e.g., 'attitude_Friendly', 'completion_time', 'spread') and firm attributes ('realized_return', 'base_equity_value') identified as salient through data-driven feature selection.

1.3 Research Aim and Research Questions

This research aims to critically evaluate the efficacy of TabPFN, a novel transformer-based architecture, in enhancing predictive accuracy and yielding interpretable insights for M&A outcomes compared to established econometric models. Specifically, this study investigates whether TabPFN's capacity to model complex, non-linear data regularities, coupled with SHAP based explainability, offers a demonstrable improvement in forecasting M&A success and provides more nuanced, actionable intelligence for strategic financial decision-making, particularly within dynamic transitional economic environments.

To achieve this aim, the research addresses the following specific questions:

- **RQ1:** How does the predictive discrimination (measured by AUC-ROC and related classification metrics) and probability calibration of the Tabular Prior-informed Bayesian Neural Network (TabPFN) model compare against benchmark tuned models, specifically Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) in forecasting M&A success on 4,400 historical U.S. transactions with announcement dates from announcement dates ranging from January 12, 2000, to July 11, 2024, and corresponding exit dates (deal completion or termination) extending from February 14, 2000, to December 26, 2024?
- **RQ2:** What are the primary determinants of M&A success as identified by SHAP values applied to the trained TabPFN and the benchmark models (LR, RF, GB, SVM, MLP)? How do the identified features (such as 'attitude_Friendly', 'realized_return', 'completion_time'), their importance rankings, and any discernible non-linear effects contrast across these models?
- **RQ3:** Based on the comparative performance (RQ1) and the nature of the explanatory insights generated (RQ2), what is the potential value and limitations of deploying an explainable TabPFN model for M&A due diligence and strategic risk assessment, particularly considering the information needs of stakeholders operating within the Transition, Innovation, and Sustainability Environments (TISE) framework?

1.4 Guiding Conceptual Lenses

This thesis uses two main complementary theoretical lenses: Complex Adaptive Systems (CAS) theory and Socio-technical Innovation. Both theories are informed by the Multi-Level Perspective (MLP) to frame the Merger and Acquisition (M&A) forecasting problem. CAS theory conceptualizes M&A environments as evolving systems composed of interacting, adaptive firms whose behavior is shaped by feedback loops, path dependence, and emergent dynamics. In such settings, non-linear relationships and sensitivity to initial conditions render traditional linear models inadequate, thereby justifying the application of AI models capable of learning from complex interdependencies and adaptive behaviors (Holland, 2006); (Uhl-Bien & Arena, 2018).

Meanwhile, the Socio-technical lens emphasizes that AI forecasting is not merely a technical innovation but is embedded within broader institutional and cultural contexts, particularly in finance where norms, routines, and cognitive framing shape tool adoption. Effective integration of AI into M&A practice requires alignment between technological capabilities and organizational interpretability, trust, and risk governance structures (Geels, 2004); (Bijker, Hughes, & Pinch, 1987). From this perspective, explainability becomes not only a functional necessity but also a socio-organizational condition for adoption.

MLP further contextualizes this process by positioning AI as a 'niche' innovation, developed and refined in protected spaces, that challenges and potentially disrupts incumbent financial forecasting regimes under the pressure of macro-level economic transitions such as digitization, climate finance, and geopolitical volatility (Geels & Schot, 2007). It links micro-level AI experimentation with meso-level organizational practices and macro-level systemic changes, offering a dynamic, multi-scalar view of technological integration in finance. These perspectives together provide a robust, systemic understanding of M&A prediction and illuminate the multifaceted challenges of integrating explainable AI into financial decision-making. They underscore that success hinges not only on predictive performance but also on the institutional, epistemological, and cultural readiness of financial systems to accommodate paradigm-shifting tools.

The thesis structure follows a logical progression: Chapter 2 Literature Review reviews literature on M&A success factors, predictive modeling, and AI interpretability. Chapter 3 Methods outlines the research methodology, including data processing and evaluation metrics. Chapter 4 Results presents empirical findings, while Chapter 5 Discussion interprets them in light of existing theories and discusses implications. Finally, Chapter 6 Conclusion synthesizes and suggests directions for future research.

2.1 M&A Success Constructs and Measurement Heterogeneity

The construct of "M&A success" is among the most contested and multifaceted in corporate finance and strategic management, reflecting both the theoretical diversity within the field and the practical complexity of real-world transactions (Bauer & Matzler, 2014); (Renneboog & Vansteenkiste, 2019); (Zollo & Meier, 2008). At its core, the challenge lies in reconciling the often divergent perspectives and objectives of various stakeholders: shareholders, managers, employees, customers, regulators, and broader society. These stakeholders may define "success" according to distinct, sometimes conflicting, criteria (King et al., 2004). This conceptual and operational heterogeneity is not merely an academic quibble; it carries profound implications for the design, comparability, and generalizability of predictive models, including AI-driven approaches, and for the validity of empirical findings across diverse M&A contexts (Zollo & Meier, 2008).

Historically, the predominant paradigm in financial economics has operationalized M&A success via short-term shareholder value creation, typically quantified by cumulative abnormal returns (CARs) surrounding the deal announcement (Renneboog & Vansteenkiste, 2019); (Zollo & Meier, 2008). This approach, while offering apparent objectivity, market-based validation, and alignment with shareholder-centric theories, faces substantial criticism. Notably, CARs primarily reflect market expectations and investor sentiment at a specific point in time, rather than realized long-term value, which often hinges on the protracted and complex process of post-merger integration (Bauer & Matzler, 2014); (Steigenberger, 2017). The utility of CARs is further limited by their sensitivity to event window specification, confounding events, and the inherent difficulty in isolating the M&A effect from broader market movements, potentially capturing ephemeral reactions rather than fundamental value shifts (Renneboog & Vansteenkiste, 2019). Zollo and Meier (2008) underscore that such short-term financial metrics often fail to correlate with measures of operational or strategic success, highlighting a disconnect between immediate market reactions and the underlying value creation processes (Zollo & Meier, 2008).

In response to these limitations, management and strategy scholars have advocated for and developed alternative constructs that emphasize long-term value creation, operational integration, synergy realization, and broader stakeholder satisfaction (Bauer & Matzler, 2014); (Steigenberger, 2017); (Gomes et al., 2013). These constructs are often multidimensional, incorporating both quantitative indicators (e.g., post-merger profitability, cost savings, market share growth, innovation output) and qualitative assessments (e.g., cultural fit, leadership continuity, employee retention, customer satisfaction) (Bauer & Matzler, 2014); (Zollo & Meier, 2008). For instance, Bauer and Matzler (2014) integrate strategic complementarity, cultural fit, and the degree and speed of integration as key antecedents, measuring success through managerial perceptions of objective and subjective outcomes. Similarly, Steigenberger's (2017) review of M&A integration literature highlights the critical role of structural and communication-based interventions in achieving integration success. The DeLone and McLean IS Success Model, adapted to M&A, suggests success as a function of strategic fit, operational synergy realization, and stakeholder value preservation (Alaranta & Henningsson, 2008). Studies employing multi-objective optimization or comprehensive managerial scales often report greater explanatory power and practical relevance, even if predictive comparisons with financial models are complex (Bauer & Matzler, 2014); (Teerikangas & Thanos, 2018).

The operationalization of these broader dimensions introduces challenges such as subjectivity, measurement error, and cross-study comparability due to metric diversity (Renneboog & Vansteenkiste, 2019). Zollo and Meier (2008) propose a structured framework linking M&A performance at task, transaction, and firm levels over varying time horizons, showing that integration processes affect acquisition outcomes, which then impact long-term firm performance, mediated by customer and employee retention (Zollo & Meier, 2008).

The implications for machine learning and AI-driven prediction are significant. Models trained on narrowly defined outcomes like short-term CARs risk capturing superficial correlations that don't generalize to long-term value creation contexts (Zollo & Meier, 2008); (Renneboog & Vansteenkiste, 2019). Meanwhile, more holistic models struggle with sparse, noisy data, overfitting, and complex feature engineering demands (Hollmann et al., 2022). Additionally, class imbalance in "successful" M&As distorts model training and evaluation unless addressed with care (Wang & Moin, 2022). Mitigating these challenges demands both methodological advances (e.g., ensemble learning, multi-label classification) and a clearly articulated, context-sensitive definition of M&A success (Steigenberger, 2017). The metric chosen ultimately defines the research task, shapes interpretation, and determines a model's usefulness in settings like the Transition, Innovation, and Sustainability Environments (TISE), where non-financial goals often dominate (Muehlmann et al., 2022; Satalkina et al., 2021, 2022).

2.2 Limitations of Econometric Baselines

Traditional econometric models, particularly logistic regression (LR) for binary outcome prediction (Salehi et al., 2019; Zhao et al., 2025) and time-series models like ARIMA for analyzing temporal patterns (Town, 1992) have long served as methodological benchmarks in the empirical analysis of M&A outcomes. These models are valued for their well-understood statistical properties, analytical transparency, and ease of implementation. Importantly, their parameters support direct inference regarding marginal effects under specific distributional and functional form assumptions (Mullainathan & Spiess, 2017). However, the increasing structural complexity of financial markets, the high dimensionality of modern M&A datasets, and the non-linear, interactive nature of many deal success drivers expose critical limitations in these traditional approaches, thereby motivating the adoption of more flexible and expressive machine learning models.

A foundational limitation of standard econometric techniques, such as LR, lies in their reliance on pre-specified functional forms, most often linear or log-linear relationships between predictor variables and the transformed outcome variable (McFadden, 1980; Zhao et al., 2025). Such constraints hinder their ability to accurately capture non-monotonic effects or complex interaction structures that often characterize M&A dynamics (Cigola & Modesti, 2008; Town, 1992). For instance, the relationship between a firm's leverage and deal completion likelihood may exhibit U-shaped characteristics, which standard LR would fail to approximate without explicitly engineered non-linear terms. Additionally, models like the multinomial logit, an extension of LR, suffer from the Independence of Irrelevant Alternatives (IIA) property, which assumes that the relative probabilities between two outcomes remain unaffected by the presence or absence of other alternatives an assumption commonly violated in multifaceted strategic decisions (McFadden, 1980).

High-dimensionality poses another formidable challenge. When the number of predictors (p) approaches or exceeds the number of observations (n), maximum likelihood estimation in LR becomes unstable or even undefined (Salehi et al., 2019). Although regularized logistic regression (RLR) introduces L1 or L2 penalties to improve estimation in such regimes, its success depends on appropriate tuning and model diagnostics, particularly under multicollinearity and interaction-heavy settings. Traditional feature selection techniques, such as stepwise regression, rely heavily on researcher discretion or simplistic criteria and may fail to identify complex or latent predictor structures (Zhao et al., 2025).

These shortcomings underscore the growing relevance of machine learning models, such as Random Forests (RF), Gradient Boosting (GB), Support Vector Machines (SVM), and Multilayer Perceptrons (MLP), which are better equipped to model nonlinearities, interactions, and high-order dependencies. RF and GB, as ensemble tree-based methods, offer robustness to multicollinearity and implicit modeling of feature interactions without requiring a priori specification (Breiman, 2001; Natekin & Knoll, 2013). SVMs, leveraging kernel-based transformations, are particularly suited for high-dimensional spaces but often suffer from scalability issues in large datasets (Cortes & Vapnik, 1995). MLPs, representing a class of deep learning models, can approximate arbitrary functions given sufficient data and architecture depth but demand careful tuning and regularization to avoid overfitting (Hornik et al., 1989). These machine learning models, in contrast to traditional LR, can also more effectively handle imbalanced datasets through integrated strategies such as class weighting, sampling techniques, and cost-sensitive loss functions (Bi & Zhang, 2021).

Another limitation of econometric models is their reliance on stationarity and time-invariant parameters, assumptions that are often violated in dynamic financial environments characterized by structural breaks, regulatory shifts, or market sentiment cycles (Town, 1992; Cigola & Modesti, 2008; Sezer et al., 2020). In M&A contexts, this is particularly salient given the episodic nature of "merger waves" and sector-specific shocks, which necessitate adaptive and temporally responsive modeling strategies.

Moreover, the application of traditional models is limited to structured, numerical data, constraining their ability to incorporate unstructured inputs such as textual disclosures, sentiment from financial news, or social media signals. In contrast, machine learning techniques, particularly when integrated with natural language processing (NLP), can parse such complex data sources to enhance predictive power (Sawhney et al., 2021; Vinocur et al., 2023).

Finally, the interpretability–predictive performance trade-off remains a central concern. Econometric models prioritize interpretability and causal inference, often at the expense of predictive accuracy, especially under misspecified functional assumptions (Mullainathan & Spiess, 2017). Conversely, machine learning models prioritize generalization to unseen data, often yielding superior performance in forecasting tasks but requiring supplementary tools for interpretability, such as SHAP values or LIME (Chatzis et al., 2018; Zhao et al., 2025). This epistemological divergence between parameter-driven causal inference and prediction-centric learning underscores the importance of model choice in applied M&A research. Where the research goal is explanatory, LR may suffice; however, for high-fidelity prediction under real-world complexity, more adaptive models such as RF, GB, SVM, MLP, and TabPFN offer compelling advantages.

2.3 Machine Learning in Corporate-Event Prediction

The inherent complexities in defining and measuring M&A success (Section 2.1) and the acknowledged limitations of traditional econometric baselines in capturing the multifaceted, non-linear, high-dimensional nature of M&A phenomena (Section 2.2) have spurred significant interest in the application of Machine Learning (ML) techniques for corporate-event prediction. ML, a subfield of artificial intelligence, shifts the focus from explanatory modeling typical of econometrics toward predictive modeling that optimizes out-of-sample forecasting accuracy (Mullainathan & Spiess, 2017). This review synthesizes the literature on ML applications in M&A, evaluates its potential to address the challenges outlined above, and motivates the primary model adopted in this thesis.

ML algorithms learn complex patterns and relationships from data without being explicitly programmed for specific functional forms (Bishop, 1994; Goodfellow et al., 2016). This capacity is especially relevant for M&A forecasting, where outcomes depend on a confluence of interacting factors spanning financial, strategic, operational, and behavioral domains (Routhu et al., 2023; Zhao et al., 2025). Building an ML model for M&A typically involves defining the prediction problem (for example, deal completion, post-merger synergy, or financial distress), preparing and engineering features to handle data idiosyncrasies, selecting appropriate algorithms, training the model, and performing rigorous validation (Lukander, O., 2025).

A substantial body of research has focused on tree-based ensemble methods such as Random Forests (RF) (Breiman, 2001) and Gradient Boosting Decision Trees (GBDTs) (Natekin & Knoll, 2013), including optimized implementations such as XGBoost (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017). These methods are widely recognized for their robustness against overfitting relative to single decision trees, capacity to model complex nonlinear interactions and mixed-type features, and provision of feature-importance metrics that support interpretability (Breiman, 2001); (Natekin & Knoll, 2013).

These ensemble models are frequently compared to other machine learning algorithms such as logistic regression (LR), support vector machines (SVM), and multilayer perceptrons (MLP). LR remains a common baseline due to its simplicity and interpretability, although it often underperforms in capturing nonlinearities (Hosmer et al., 2013). SVMs provide robust performance in high-dimensional settings with margin-based optimization but are computationally expensive for large-scale problems (Cortes & Vapnik, 1995). MLPs, as a class of feedforward artificial neural networks, are effective at capturing complex patterns but require careful tuning and large datasets to generalize well (Hornik et al., 1989).

In empirical applications, Zhao et al. (2025) found that RF outperformed traditional LR models in predicting M&A activity in the Chinese market, with acquisition experience and ownership structure emerging as dominant predictors. Lukander (2025) similarly identifies tree-based ensembles, particularly RF and GBDTs, as prevalent in M&A outcome modeling, noting their high predictive precision. A comparative study by Chatzis et al. (2018) on forecasting stock market crises, a related class of rare but impactful financial events, also underscores the superiority of RF and XGBoost over baseline methods, including LR and SVM, in such high-stakes classification tasks (Chatzis et al., 2018).

Neural networks, particularly deep learning architectures, constitute another important avenue of ML application in finance and increasingly in M&A (Bishop, 1994; Sezer et al., 2020). These models can learn hierarchical feature representations from raw data, making them well-suited to complex, high-dimensional inputs including unstructured text or multimodal sources (Goodfellow et al., 2016). Sawhney et al. (2021) introduced a multimodal approach using conference call transcripts and audio (the M3A dataset) to forecast financial risk, demonstrating the value of data beyond traditional financial statements. Vinocur et al. (2023) applied Natural Language Processing to annual reports to quantify M&A capability and link it to long-term performance, operationalizing strategic management constructs through ML.

Nevertheless, challenges persist. Data scarcity and quality issues arise especially for private-firm deals or niche transaction types (Lukander, 2025; Routhu et al., 2023). The opacity of complex ML models, notably deep neural networks, creates regulatory and managerial concerns around interpretability. Class imbalance between few failed deals versus many successful ones requires specialized training and evaluation methods to prevent biased predictions (Bi & Zhang, 2021). Addressing these limitations entails robust validation strategies such as k-fold cross-validation and out-of-time testing, and evaluation metrics sensitive to imbalance, like the area under the precision-recall curve (Lukander, 2025; Zhao et al., 2025). The integration of explainable AI techniques is crucial for aligning model decisions with financial theory and domain expertise (Routhu et al., 2023; Vinocur et al., 2023).

In sum, ML provides powerful tools for M&A corporate-event prediction, with tree-based ensembles and foundation models offering substantial gains in predictive accuracy over traditional econometric approaches, especially in high-dimensional, non-linear settings. The capacity to leverage novel data sources text, audio, synthetic series and to quantify feature importance and uncertainty makes ML methods a promising frontier. Yet effective deployment demands careful attention to data quality, validation rigor, interpretability, and computational feasibility within the unique context of M&A forecasting.

2.4 Transformers for Tabular Inference

The transformative impact of the Transformer architecture (Vaswani et al., 2017) on natural language processing (NLP) and computer vision has catalyzed significant interest in its application to tabular data, a ubiquitous format in finance, healthcare, and business analytics (Badaro et al., 2023). Unlike sequential text or gridded image data, tabular datasets present unique challenges: they often lack inherent sequential or spatial structure, feature heterogeneous data types (categorical and continuous), and may not possess strong local structural priors that attention mechanisms readily exploit (Hollmann et al., 2023; Müller et al., 2022). This section critically examines the evolution of Transformer-based models for tabular inference, from initial architectural adaptations to the sophisticated meta-learning paradigm embodied by the Tabular Prior-informed Bayesian Neural Network (TabPFN), which is central to this thesis. The discussion will also incorporate the theoretical underpinnings of the feature engineering strategies employed in this study.

Early efforts to adapt Transformers for tabular data focused on modifying input representations and attention mechanisms. TabTransformer, for instance, processes categorical features by learning contextual embeddings via self-attention, while continuous features are handled by a separate multi-layer perceptron (MLP) (Huang et al., 2020). FT-Transformer (Feature Tokenizer Transformer) extended this by tokenizing all features (both categorical and continuous) into embeddings and applying a stack of Transformer layers to capture inter-feature interactions (Gorishniy et al., 2021). These models demonstrated competitive performance against traditional MLPs and, in some cases, approached the efficacy of gradient-boosted decision trees (GBDTs), albeit often requiring substantial datasets and extensive hyperparameter tuning (Gorishniy et al., 2021). A broader survey by Badaro et al. (2023) details various strategies for input processing (e.g., serialization, filtering) and encoder adaptations (e.g., specialized positional embeddings, modified attention patterns) in this context.

A significant paradigm shift occurred with the introduction of Prior-Data Fitted Networks (PFNs), which leverage meta-learning principles (Finn et al., 2017; Snell et al., 2017) to train Transformers for a novel objective: approximating Bayesian inference (Müller et al., 2022). Instead of being trained on a single task, PFNs are meta-trained on a multitude of synthetic supervised learning tasks sampled from a carefully constructed prior distribution over data-generating processes. The core mathematical logic, as elucidated by Müller et al. (2022), is to learn a function $q\theta$ that approximates the true Bayesian Posterior Predictive Distribution (PPD), $P(y * | x *, Dtrain)$ for a new test point $x*$ given a training dataset $Dtrain = \{(xi, yi)\}i = 1N$. The PFN, parameterized by θ, takes the training set $Dtrain$ and the test input $x *$ as a single, set-valued sequence and, in a single forward pass (in-context learning), outputs the PPD. The training objective is to minimize the expected negative log-likelihood (Prior-Data NLL) over the distribution of synthetic tasks:

$$LPFN =- E((xtest, ytest) \cup Dtrain) \sim p(D)[logq\theta(ytest|xtest, Dtrain)]$$

where $p(D)$is the prior distribution over datasets. This formulation allows the PFN to effectively "learn to learn," internalizing inductive biases from the prior that enable rapid adaptation to new, small tabular datasets without explicit retraining or hyperparameter tuning. The architecture typically employs a Transformer encoder without standard positional encodings to maintain permutation invariance with respect to the training examples in $Dtrain$ (Müller et al., 2022). For regression tasks, PFNs may use a discretized continuous (Riemann) distribution head to model the output.

TabPFN, the specific instantiation used in this thesis, applies this PFN framework to tabular classification, having been pre-trained on millions of synthetic datasets generated from priors incorporating Bayesian Neural Networks (BNNs) and Structural Causal Models (SCMs) (Hollmann et al., 2023; Müller et al., 2022). This pre-training endows TabPFN with a strong inductive bias towards simplicity and causal structures, making it particularly effective for small sample sizes (Hollmann et al., 2023; McCarter, C., 2022). Its ability to perform inference in a single forward pass results in exceptionally fast predictions, often outperforming complex AutoML pipelines in speed while maintaining competitive accuracy (Hollmann et al., 2023). The more recent TabPFN v2 extends these capabilities, notably by incorporating "randomized feature tokens" to better handle heterogeneous feature spaces and improve generalization (Ye et al., 2025). This mechanism involves projecting each raw attribute value into a shared k-dimensional space and then adding a unique, random k-dimensional perturbation for each attribute, allowing the model to differentiate features without pre-defined semantics (Hollmann et al., 2025, supplementary material; Ye et al., 2025). This is a departure from earlier models like TabTransformer or

FT-Transformer that learn specific embeddings for each feature. While this thesis primarily utilizes the established TabPFN (v1 based on Müller et al., 2022 and Hollmann et al., 2023), the conceptual advancements in handling feature heterogeneity in v2 are noted. Sarwo and Prabowo (2023) provide an application-specific example where TabPFN, alongside XGBoost and LightGBM, was evaluated for classification optimization, achieving improved Balanced Log Loss, underscoring its utility in practical scenarios. Furthermore, TabPFN's versatility has been demonstrated by adapting it to time series forecasting through featurization of timestamps (Hoo et al., 2025) and exploring its scalability using techniques like dataset distillation.

The efficacy of TabPFN, like any advanced tabular model, is intrinsically linked to the quality and relevance of its input features. In this thesis, the feature engineering strategy is anchored in established financial and strategic management theories to ensure that the selected predictors are not only statistically informative but also theoretically coherent. This dual emphasis grounds the modeling process in a framework that promotes both empirical robustness and interpretability, which is particularly critical for high-stakes forecasting tasks such as M&A outcome prediction.

The predictive efficacy of TabPFN hinges on the relevance of its inputs. Accordingly, the present study begins with a 36-variable set extracted from the “Model Training Data.xlsx – full data.csv” file, reflecting both classical finance theory and strategic management considerations. Deal-level dimensions include base_equity_value (deal size), day_premium and spread (valuation gaps between offer and market price), payment_method, completion_time, and historical_avg_completion_time, while the transactional climate is captured by attitude, one-hot encoded as Friendly, Neutral, or Hostile. Firm-specific variables comprise acquirer_marketcap, a binary friendly flag, high_quality (issuer-level quality), and realized_return where available. External context is proxied by hhi_score, and procedural complexity by shareholder_vote requirements. Industry heterogeneity is encoded through target_naics_code, expanded into binary indicators for the major NAICS sectors represented in the sample.

To retain only the most informative signals and to avoid overfitting, a Random Forest importance analysis (Breiman, 2001) identifies the ten most salient predictors of post-deal success: attitude_Friendly, realized_return, friendly, completion_time, attitude_Hostile, spread, attitude_Neutral, historical_avg_completion_time, base_equity_value, and day_premium. These features are standardized via z-scores to ensure numerical commensurability. The resulting parsimonious yet theoretically grounded feature subset serves as the sole input to TabPFN and all benchmark models, balancing explanatory power with computational efficiency while preserving interpretability for subsequent managerial analysis.

In summary, TabPFN stands as a sophisticated application of Transformer architectures to tabular data, leveraging Bayesian meta-learning to deliver rapid, accurate, and well-calibrated predictions, especially in data-scarce settings. Its unique in-context learning mechanism, when combined with theoretically informed feature engineering, establishes a robust and interpretable framework for M&A outcome forecasting. This approach addresses critical limitations of both traditional econometric models and prior deep learning approaches.

2.5 Explainable AI in Regulated Finance

The increasing adoption of sophisticated Artificial Intelligence (AI) and Machine Learning (ML) models across the financial sector spanning domains such as credit risk, fraud detection, algorithmic trading, and merger and acquisition (M&A) analysis offers tremendous promise for enhanced accuracy and efficiency (Badaro et al., 2023; Zhou et al., 2024). However, this evolution also exposes institutions to significant risks around transparency, accountability, and regulatory compliance. In contrast to classical econometric models whose structure and assumptions are explicitly defined, modern AI systems, particularly deep learning architectures like Transformers (used in models such as TabPFN), often behave as complex nonlinear function approximators (Hollmann et al., 2023; Schnurr et al., n.d.). Ensemble models such as Random Forests and Gradient Boosting Machines (GBMs), while not deep learning, also present interpretability challenges (Antar & Tayachi, 2025; Tan et al., 2023). These models frequently operate as “black boxes,” where their internal logic is not easily interpretable, even by their developers (Murdoch et al., 2019; Yeo et al., 2025). This opacity is especially problematic in finance, a domain governed by strict regulatory scrutiny and high thresholds for auditability, fairness, and model risk management (Fritz-Morgenthal et al., 2022).

The inability to clearly explain model outputs undermines trust, exposes institutions to operational and reputational risks, and potentially violates laws related to anti-discrimination, creditworthiness evaluation, or algorithmic decision-making. Regulatory bodies have increasingly acknowledged that this lack of transparency can amplify biases, obscure systemic vulnerabilities, and obstruct the governance of decision-critical systems (Laux et al., 2024). The European Union’s Artificial Intelligence Act (EU AI Act), for instance, codifies this shift towards demanding transparency, particularly for high-risk AI systems commonly found in finance (Černevičienė & Kabašinskas, 2024; Kusche, I., 2024).

Article 13 of the EU AI Act mandates that AI systems be transparent enough to allow users to interpret outputs and make appropriate use, while Article 14 requires models to permit effective human oversight (European Commission, 2021b, as cited in Laux et al., 2024). Regulatory bodies like Germany's BaFin and the European Securities and Markets Authority (ESMA) have also issued guidance emphasizing the need for XAI tools capable of generating stakeholder-specific explanations (Fritz-Morgenthal et al., 2022).

In response to these technical, ethical, and regulatory imperatives, the field of Explainable AI (XAI) has emerged. XAI comprises methods designed to clarify how complex AI systems generate predictions or classifications, enabling both technical and non-technical stakeholders to assess model rationale, fairness, and robustness (Arrieta et al., 2020, as cited in Yeo et al., 2025). In the financial domain (FinXAI), techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have become prominent (Lundberg & Lee, 2017; Ribeiro et al., 2016). SHAP values, rooted in cooperative game theory, quantify the marginal contribution of each input feature to a specific prediction, offering both local (instance-level) and global (model-level) interpretations (Lundberg & Lee, 2017). SHAP has been applied to various financial tasks, including M&A outcome prediction (Zhou et al., 2024), credit scoring (Xia et al., 2019), financial stability assessment (Tan et al., 2023), and derivatives pricing (Davis et al., 2023). LIME, on the other hand, builds sparse linear models around individual predictions to explain the behavior of any black-box classifier locally (Ribeiro et al., 2016, as cited in Murdoch et al., 2019). Its utility has been demonstrated in areas such as credit default prediction (Antar & Tayachi, 2025) and understanding complex text-based financial models (Yeo et al., 2025).

While these XAI methods provide crucial bridges between advanced model architectures and the interpretability demanded by stakeholders, they are not without limitations. The computation of exact SHAP values can be intensive for complex models and large datasets (Lundberg & Lee, 2017), although approximations like KernelSHAP or TreeSHAP offer more efficient alternatives (Davis et al., 2023). LIME's explanations are local and may not always generalize well or remain stable across slight perturbations (Murdoch et al., 2019). Furthermore, both SHAP and LIME, in their common implementations, often assume feature independence, which can lead to misleading interpretations in financial datasets where multicollinearity is prevalent (Aas et al., 2021, as cited in thesis draft). Critiques from the XAI literature also highlight that XAI tools are themselves models with their own assumptions and limitations, demanding rigorous validation (Fritz-Morgenthal et al., 2022; Sabbatini & Manganaro, 2023).

The choice of background dataset for SHAP, for instance, can significantly influence attributions (Covert & Lee, 2020, as cited in thesis draft).

The application of TabPFN in this thesis, a transformer-based model leveraging Bayesian inference (Hollmann et al., 2023; Schnurr et al., n.d.), inherently operates as a "black box" despite its sophisticated prior-informed architecture. Thus, post hoc explanation methods like SHAP are indispensable for dissecting its predictions, particularly in M&A forecasting where understanding the drivers of success or failure (e.g., deal characteristics, acquirer financial health, innovation metrics) is critical for strategic decision-making and regulatory compliance (Zhou et al., 2024). The ability to generate human-interpretable explanations from TabPFN's outputs allows analysts to validate model logic against financial theory, identify potential biases, and communicate findings to diverse stakeholders, from investment committees to regulatory bodies. This aligns with the overarching goal of XAI in regulated finance: to ensure that as AI systems become more powerful, they also become more transparent, accountable, and trustworthy, thereby fostering responsible innovation (Yeo et al., 2025; Fritz-Morgenthal et al., 2022). The consistent finding across multiple studies (e.g., Antar & Tayachi, 2025; Xia et al., 2019; Tan et al., 2023) is that explainability is not merely a technical add-on but a fundamental component of deploying AI in high-stakes, compliance-intensive domains.

2.6 Integrative Gap Statement and Resulting Hypotheses

The literature review (Sections 2.1–2.5) has explained a multifaceted and persistent set of challenges at the intersection of M&A outcome prediction, the limitations of traditional econometric modeling, the capabilities of advanced AI architectures, and the urgent demands for explainability in regulated financial environments. A critical synthesis of this literature, now augmented by recent works in FinXAI (Černevičienė & Kabašinskas, 2024; Yeo et al., 2025) and specific applications of machine learning to M&A and related financial forecasting (Zhou et al., 2024; Antar & Tayachi, 2025; Tan et al., 2023), reveals an integrative gap. This gap centers on the insufficient empirical validation of novel, explainable AI models like TabPFN within the high-stakes context of M&A forecasting, especially regarding their capacity to enhance predictive accuracy while providing transparent, actionable insights that comply with evolving regulatory frameworks such as the EU AI Act (Fritz-Morgenthal et al., 2022; Laux et al., 2024). Several interconnected issues contribute to this overarching gap:

First, there is persistent methodological heterogeneity and clear limitations in M&A success forecasting. As established in Sections 2.1 and 2.2, the definition and measurement of M&A success remain fragmented (Zollo & Meier, 2008), and traditional econometric models such as logistic regression struggle to capture the non-linearities, high-dimensional interactions, and dynamic nature of M&A datasets (Mullainathan & Spiess, 2017). These models often fail to account for the complex interplay of deal, firm, and environmental characteristics that ultimately drive M&A outcomes (Zhou et al., 2024).

Second, while advanced AI models, including machine learning techniques like tree-based ensembles (Tan et al., 2023; Antar & Tayachi, 2025; Xia et al., 2019) and transformer architectures such as TabPFN (Hollmann et al., 2023; Müller et al., 2022; Schnurr et al., n.d.), have demonstrated the ability to model these complexities with greater predictive power, their “black-box” nature poses a significant barrier to adoption in regulated financial domains (Arrieta et al., 2020, as cited in Yeo et al., 2025; Fritz-Morgenthal et al., 2022). The internal logic of these models is often not readily interpretable, even by their developers (Murdoch et al., 2019).

Third, there is an increasing regulatory imperative for explainable AI in finance. The regulatory landscape, shaped by the EU AI Act (European Commission, 2021b) and guidance from bodies such as BaFin and ESMA, now mandates transparency, auditability, and human oversight for high-risk AI systems (Section 2.5; Laux et al., 2024; Kusche, I. 2024). While explainability methods such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) seek to bridge this gap, their application, stability, and practical utility especially when integrated with advanced models like TabPFN for M&A forecasting still require rigorous empirical validation (Černevičienė & Kabašinskas, 2024). Additional issues such as the computational cost of explainability methods, sensitivity to background data (Covert & Lee, 2020, as cited in thesis draft), and the assumption of feature independence can limit their reliability and effectiveness (Aas et al., 2021).

Another challenge is the lack of integrated predictive and explanatory evaluation in M&A contexts. Few studies, particularly within M&A forecasting, have systematically compared advanced AI models like TabPFN to traditional baselines while also evaluating the depth and actionability of their explainability outputs. While Zhou et al. (2024) provide a relevant benchmark using gradient boosting machines and SHAP for M&A in China, the unique meta-learning approach and Bayesian inference capabilities of TabPFN warrant dedicated investigation. The practical value proposition of such models rests not only on marginal gains in predictive accuracy but critically on their ability to deliver trustworthy, interpretable, and decision-relevant insights.

The imperative for the TISE framework (Transition, Innovation, and Sustainability Environments) further amplifies these needs. Strategic resource allocation within

TISE demands forecasting tools that do more than just predict; they must support nuanced understanding of complex, dynamic systems and facilitate decision-making aligned with stakeholder interests under uncertainty. This requires models that are robust, interpretable, and capable of integration into broader governance and strategic planning processes (Muehlmann et al., 2022; Satalkina et al., 2021, 2022).

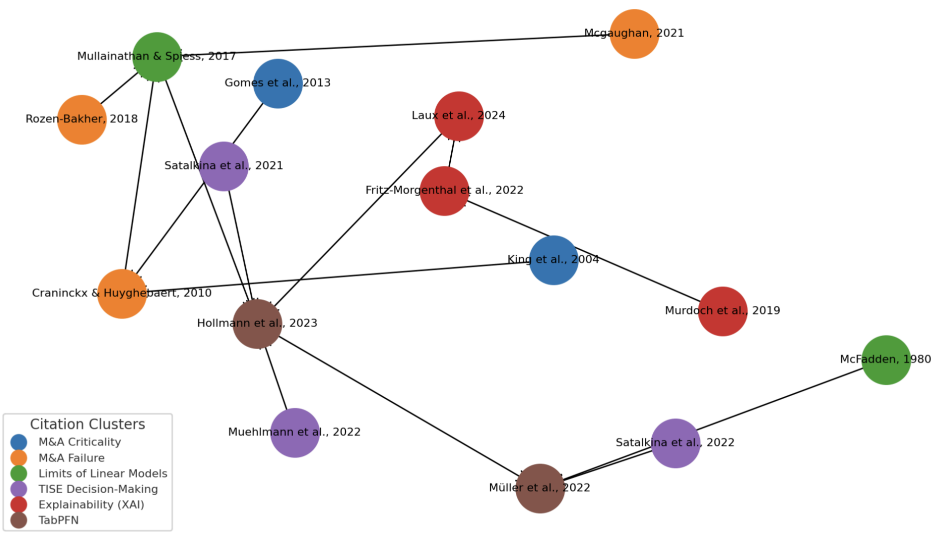
This confluence of issues underscores the central research problem: the need for a rigorous, empirically grounded assessment of an advanced, explainable AI model (TabPFN with SHAP) for M&A outcome forecasting. Specifically, this evaluation must address predictive superiority, the nature of model explanations, and practical utility within the demanding TISE framework. Accordingly, this thesis formulates the following hypotheses to address these gaps:

- **H1:** The Prior-informed Bayesian Neural Network (TabPFN) model will demonstrate significantly higher discriminative performance, as measured by the Area Under the Receiver Operating Characteristic Curve (AUC) and F1-score, and superior probability calibration, as measured by Brier score and Expected Calibration Error (ECE), compared to the traditional logistic regression baseline model in forecasting M&A success or failure on the selected dataset. (Rationale: Drawing from TabPFN's established efficacy in capturing complex patterns in tabular data and its superior performance in various machine learning benchmarks (Hollmann et al., 2023; Schnurr et al., n.d.), it is expected to outperform the linear and additive assumptions of logistic regression in the multifaceted domain of M&A outcomes.)
- **H2:** Explainable AI analysis using SHAP, when applied to the trained TabPFN model, will identify influential features driving M&A outcome predictions and potentially reveal non-linear effects or feature interactions that are not readily apparent in the logistic regression baseline. (Rationale: The capacity of complex models like TabPFN to learn non-linear relationships, combined with SHAP's ability to decompose predictions into feature-level attributions (Lundberg & Lee, 2017), is expected to provide richer and more nuanced explanatory insights than those derivable from the linear coefficients of logistic regression, as demonstrated in applications such as credit default and financial stability forecasting (Antar & Tayachi, 2025; Tan et al., 2023).)
- **H3:** The combination of TabPFN's predictive outputs, including calibrated probabilities, and the feature importance insights derived from SHAP explanations, will provide a more valuable and actionable tool for M&A strategic decision-making compared to the predictive output of the logistic regression baseline alone, especially within the TISE framework.

(Rationale: Enhanced predictive accuracy and probability calibration (as per H1), together with transparent and granular feature-level explanations (as per H2), should result in greater decision-maker confidence, improved risk assessment, and more informed strategy formulation (Yeo et al., 2025). This synergy addresses the TISE context's demand for robust, interpretable, and trustworthy AI tools for navigating complex transitions (Muehlmann et al., 2022).)

By testing these hypotheses, this study aims to contribute to both the methodological advancement of AI in finance and its practical application in strategic corporate decision-making. The core objective is to evaluate whether high-capacity, explainable deep learning models can meet the dual demands of predictive excellence and regulatory or stakeholder transparency in the complex domain of M&A forecasting.

Figure 2.6. Main Literature Reviews Connections. Made by the Author.



3.1 Research Paradigm & Design

This research adopts a positivist paradigm, which is rooted in the philosophical assumption that reality exists independently of human perception and can be objectively measured through empirical observation (Collis & Hussey, 2014). Positivism prioritizes quantifiable data, formal hypothesis testing, and the identification of causal relationships or statistical regularities. Within this worldview, the role of the researcher is to maintain detachment, systematically test hypotheses, and draw conclusions based on observable, replicable phenomena. Such an approach is particularly appropriate for computational finance applications, where empirical performance, statistical significance, and model generalizability are paramount.

This ontological and epistemological position aligns directly with the core aim of this thesis: to evaluate and compare the predictive capabilities of two competing modeling approaches, TabPFN and regularized logistic regression, on the measurable binary outcome of M&A success or failure. The selection of Area Under the Receiver Operating Characteristic Curve (AUC), F1-score, and probability calibration metrics as evaluation criteria reinforces the commitment to empirical rigor and objective comparability. These metrics are well-established in both statistical learning and applied finance literature for assessing classifier performance, particularly when imbalanced classes and probabilistic outputs are involved (Niculescu-Mizil & Caruana, 2005); (Siblini et al., 2019).

Underpinning this paradigm is a research design rooted in predictive analytics, structured to generate and validate empirical forecasts and then compare the performance of those forecasts across different model families. The design consists of two analytically distinct but methodologically integrated stages. The predictive component involves training each model on a structured historical M&A dataset to forecast deal success using only features available ex-ante. This simulates the kind of forward-looking analysis employed in real-world M&A strategy or due diligence contexts. The comparative component then evaluates the models' outputs on a held-out test set, using consistent quantitative benchmarks to determine relative predictive power and model calibration. The use of a standardized out-of-sample test set mitigates the risk of overfitting and supports generalizable claims within the scope of the data.

Crucially, while the research remains rooted in prediction, it integrates a secondary explanatory objective that enhances its methodological sophistication and practical relevance. Specifically, explainable AI techniques such as SHAP are applied to the outputs of the TabPFN model to uncover the structure of the learned decision surface. This explanatory layer does not alter the positivist foundation, since it still relies on empirical, algorithmically derived attributions, but enriches the analysis by offering interpretable insights into why the model reaches specific classifications.

The inclusion of SHAP-based interpretation also addresses a key limitation of black-box models in finance: their limited utility in decision support settings where transparency is not merely desirable but often required. By enabling comparison of feature importances and interaction effects across models, the design permits an evidence-based assessment of not only how well models perform (RQ1/H1), but also how their internal logic aligns, or fails to align, with domain knowledge and economic theory. This integrated design reflects a growing consensus in the field that prediction and explanation need not be in opposition but can be productively synthesized to yield actionable insights in complex decision environments (Doshi-Velez & Kim, 2017).

Finally, the structure of the design supports the practical implications central to RQ3 and H3. By comparing not only output accuracy but also interpretability and strategic relevance, the research tests whether advanced, explainable ML models like TabPFN can credibly support decision-making in regulated, high-stakes financial domains. The predictive-comparative design therefore operationalizes the TISE framework by generating findings that are not only statistically valid but also practically actionable within environments defined by transition, innovation, and sustainability constraints.

3.2 Data Source, Sampling, Inclusion/Exclusion, Ethical Compliance

Consistent with the data provider's labeling and a common operationalization in predictive M&A studies, 'success' for this research is defined as deal completion, while 'failure' corresponds to deal termination. This binary outcome, while an abstraction of the multifaceted M&A success constructs discussed in Section 2.1 (e.g., Zollo & Meier, 2008), provides a clear, objectively measurable target for the classification models. The predictive modeling framework in this study was initialized with a comprehensive set of 36 features, meticulously engineered or selected from the M&A transaction data based on their established theoretical relevance to M&A outcomes within corporate finance and strategic management literature, alongside their demonstrated empirical utility in prior financial forecasting studies. This feature set was designed to provide a multifaceted representation of each transaction, encompassing critical dimensions such as valuation metrics, deal structure intricacies, process characteristics, firm-specific attributes of both acquirer and target, the prevailing market context, and detailed industry classifications.

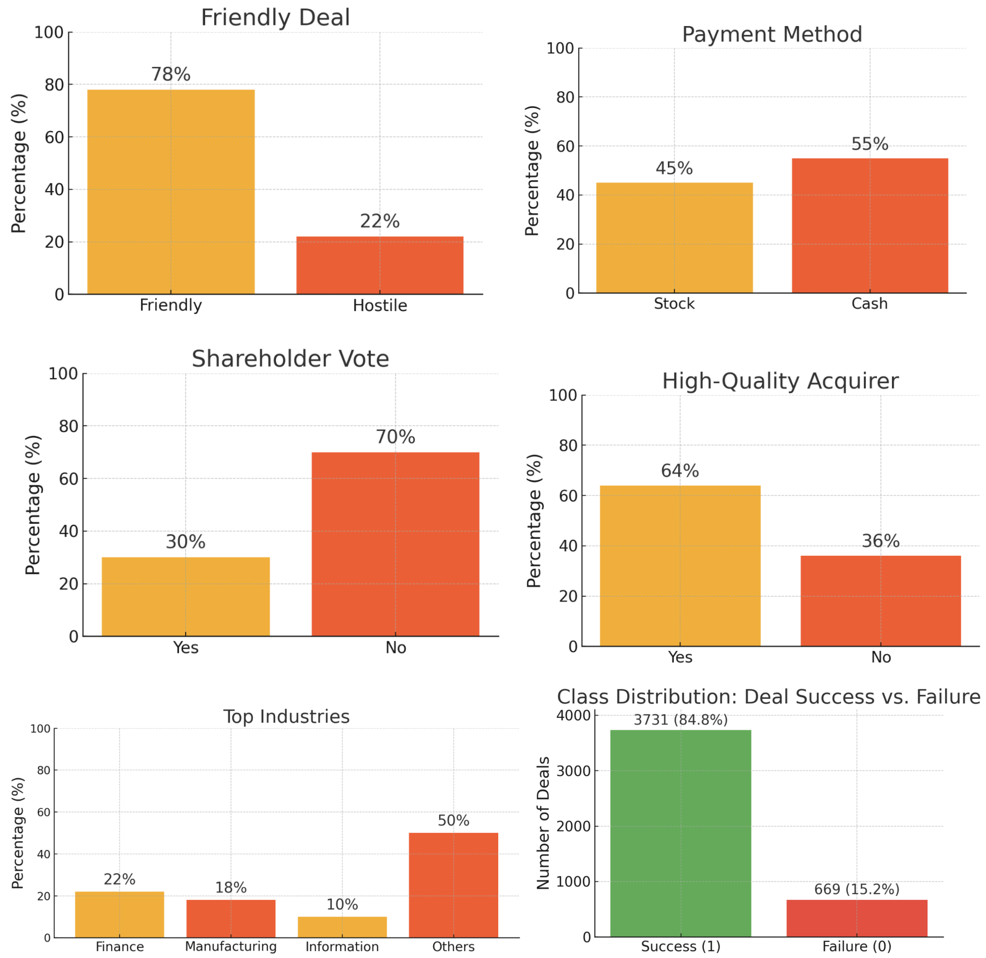
Specifically, the input variables (post one-hot encoding of the 'attitude' feature but prior to any feature selection or pruning algorithms) included: base_equity_value, a continuous variable representing the target company's equity value at announcement, scaled (e.g., via z-score standardization) to serve as a primary indicator of deal size, which itself influences integration complexity, regulatory scrutiny, and strategic motivations, potentially reflecting economies of scale or risks of "empire building" (Jensen, 1986). payment_method was a binary feature distinguishing stock-financed deals (value '1') from those primarily using cash or a mix (value '0'), crucial for its signaling effects regarding acquirer valuation perceptions (Myers & Majluf, 1984), risk allocation, and tax implications that impact deal success (Travlos, 1987). The necessity of formal shareholder approval from either party was captured by shareholder_vote (Binary), representing an additional procedural hurdle and source of uncertainty. The deal's initiation nature was decomposed via one-hot encoding of the original categorical 'attitude' variable into three binary features: attitude_Friendly ('1' for target management support, '0' otherwise), generally linked to smoother due diligence and higher success probabilities (Schwert, 2000); attitude_Hostile ('1' for target management opposition, '0' otherwise), often entailing higher premiums, costs, and integration challenges; and attitude_Neutral ('1' if the stance was unclear or uncommitted, '0' otherwise), covering unsolicited but not overtly opposed bids.

Valuation and market perception were further detailed by day_premium, a scaled continuous variable for the premium offered to target shareholders relative to the pre-announcement stock price, reflecting immediate financial incentive and acquirer's synergy valuation, though excessive premiums can signal overpayment. The acquirer_marketcap, a scaled continuous variable, denoted the acquirer's pre-deal market capitalization, proxying for its size, financial resources, market presence, and M&A experience, with implications for absorption capacity versus integration complexity. A proprietary or financially derived binary indicator, high_quality, likely proxied the perceived quality or financial strength of involved entities, hypothesized to correlate with success. Reinforcing the deal's nature, friendly (Binary) directly indicated cooperative transaction terms, likely correlated with attitude_Friendly. Short-term market reactions were captured by realized_return, a scaled continuous variable measuring abnormal returns around the announcement, reflecting immediate market assessment of value-creation potential, distinct from post-completion performance to avoid data leakage. The spread (Continuous, Scaled) measured the deal spread post-announcement but pre-completion, with wider spreads indicating higher perceived completion risk. Market structure was addressed by hhi_score (Continuous, Scaled), the Herfindahl-Hirschman Index for the target's industry, indicating market concentration which could imply less competition or attract antitrust scrutiny.

Process characteristics included completion_time (Continuous), the duration in days from announcement to exit, where longer times can signal hurdles or deteriorating fundamentals. This was contextualized by historical_avg_completion_time (Continuous), the average completion time for similar past deals, helping identify significant deviations.

The remaining 21 features consisted of binary dummy variables derived from the North American Industry Classification System (NAICS) codes of the target company, allowing for the incorporation of industry-specific effects. These included: 51 (Information, e.g., software, telecom), 53 (Real Estate and Rental and Leasing), 31 (Manufacturing, often 31-33 supersector), 42 (Wholesale Trade), 22 (Utilities, heavily regulated), 44 (Retail Trade, part of 44-45 supersector), 11 (Agriculture, Forestry, Fishing and Hunting), 52 (Finance and Insurance), 54 (Professional, Scientific, and Technical Services), 55 (Management of Companies and Enterprises), 21 (Mining, Quarrying, and Oil and Gas Extraction), 56 (Administrative and Support and Waste Management and Remediation Services), 72 (Accommodation and Food Services), 62 (Health Care and Social Assistance), 61 (Educational Services), 48 (Transportation and Warehousing, often 48-49 supersector), 23 (Construction), 45 (Retail Trade, potentially a distinct subsector from 44, further specifying retail environment), 71 (Arts, Entertainment, and Recreation), 92 (Public Administration, with unique regulatory/public interest aspects), and 81 (Other Services, a residual category). This comprehensive set of 36 features, following preprocessing steps like imputation, one-hot encoding for 'attitude', and standardization for continuous variables, constituted the initial input space for the Random Forest-based feature selection algorithm, which subsequently identified the top-10 most impactful predictors for the final predictive models, thereby enabling a nuanced examination of M&A dynamics by bridging financial metrics, deal characteristics, and rich industry contexts.

Figure 3.2. Categorical Feature Frequencies Example. Made by the Author



3.3 Feature Engineering Rationale

From the original 36 predictors the feature set was refined through a sequential, data-driven procedure. First, 2 000 randomly selected training observations were balanced with the Synthetic Minority Over-sampling Technique (SMOTE, random_state = 42) to offset class imbalance and stabilise importance estimates (Chawla et al., 2002). A Random Forest classifier with 50 trees (random_state = 42) was then fitted to this balanced sample, and Gini importances were computed for every variable (Breiman, 2001). The ten most informative predictors—attitude_Friendly, realized_return, friendly, completion_time, attitude_Hostile, spread, attitude_Neutral, historical_avg_completion_time, base_equity_value, and day_premium—were retained for modelling. Their theoretical relevance aligns with merger research: deal sentiment indicators shape information symmetry and integration prospects, with friendly approaches typically exhibiting higher success odds (Schwert, 2000); valuation and performance measures capture market expectations of synergy and deal size; temporal variables benchmark execution complexity relative to historical norms. All continuous features were subsequently standardised with z-scores to ensure numerical commensurability before final model training.

3.4 Pre-Processing Pipeline

The raw dataset was subjected to a structured preprocessing pipeline that converted categorical sentiment indicators, addressed missing data, pruned features, and standardised continuous variables before model training. First, the attitude column was one-hot encoded into three binary variables (attitude_Friendly, attitude_Hostile, attitude_Neutral) to retain deal sentiment information. Numerical features with missing values were imputed with their column medians, a robust approach against outliers, and any remaining gaps were filled with zero to maintain dimensional consistency. The full 36-variable matrix was then reduced to the ten most salient predictors via Random Forest Gini importance, calculated on a class-balanced subset produced with the Synthetic Minority Over-sampling Technique (SMOTE, random_state = 42) as detailed in Section 3.3. All retained continuous features were z-standardised using scikit-learn's StandardScaler to ensure zero mean and unit variance, facilitating convergence in support vector machines, multilayer perceptrons, logistic regression, and improving numerical stability for tree ensembles. The SMOTE-augmented and scaled dataset (Xr_p_scaled, yr) supported baseline hyperparameter tuning and trained the TabPFN model, while the original training, validation, and test partitions (X_train_p, X_val_p, X_test_p) were rescaled with the same fitted scaler to preserve out-of-sample comparability. Each transformation was implemented in Python with fixed seeds, and the corresponding script is available in merger_pipeline_full.py (lines 45–131) in the Index/ANNEXES section D.

3.5 Model Specifications

This study benchmarks the Tabular Prior-informed Bayesian Neural Network (TabPFN) against five tuned baselines: logistic regression (LR), random forest (RF), gradient boosting (GB), support vector machine (SVM), and multilayer perceptron (MLP). TabPFN, implemented via the *tabpfn* library and run on CPU, was restricted to ten seconds per fit or predict call (AUTO_MAX_SEC = 10) on the SMOTE-balanced, ten-feature training matrix Xr_p. The model employs a pre-trained transformer that meta-learns a Bayesian prior over synthetic tabular problems, allowing near-instant inference on small datasets and producing calibrated probabilistic outputs whose quality is later assessed via expected calibration error and the Brier score (Hollmann et al., 2023; see Sections 3.7 and 4.2).

All baselines were coded in scikit-learn and tuned with *RandomizedSearchCV* using two random hyperparameter draws and a two-fold stratified cross-validation scheme, optimising the area under the receiver-operating-characteristic curve with a fixed seed of 42 for reproducibility. The LR grid varied CCC ∈ {0.1, 1} under an L2 penalty and the *liblinear* solver with 500 iterations. The RF grid evaluated 50 trees, tree depth either unlimited or capped at five, and a minimum split size of two, exploiting all available CPU cores. The GB search tested 50 trees of depth three with a learning rate of 0.1. The SVM search probed CCC ∈ {0.1, 1} with an RBF kernel and γ set to *scale* while enabling probabilistic outputs. The MLP search examined a single hidden layer of 50 neurons with $\alpha = 10^{-3}$, learning_rate_init = 10^{-3}, early stopping, and a limit of 200 iterations. The best estimator from each search was refitted on the full SMOTE-balanced, standardised training data and finally evaluated on the likewise scaled hold-out test set $Xtest\_p\_scaled$. This systematic pipeline affords a rigorous comparison of TabPFN's Bayesian meta-learning approach with conventional machine-learning classifiers tuned under identical data conditions.

3.6. Architectural and Inferential Framework

The primary advanced AI model evaluated in this thesis is the Tabular Prior-informed Bayesian Neural Network (TabPFN), a novel architecture rooted in the concept of Prior-Data Fitted Networks (PFNs) that leverages meta-learning to approximate Bayesian inference for tabular classification tasks (Müller et al., 2022; Hollmann et al., 2023). Unlike traditional models trained on a single dataset for a specific task, TabPFN is pre-trained on a vast ensemble of synthetic supervised learning tasks. These tasks are sampled from a structured prior distribution over data-generating processes, allowing the model to internalize a rich set of inductive biases that mimic Bayesian reasoning. This pre-training endows TabPFN with the ability to perform "in-context learning": it can make predictions on new, small tabular datasets (typically up to ~1,000 samples and ~100 features, as used in this study after pruning) in a single forward pass, without requiring task-specific fine-tuning or extensive hyperparameter optimization.

The core mathematical objective of a PFN, and thus TabPFN, is to learn a function $T\Phi*$ (where $\Phi*$ represents the optimized parameters from meta-training) that approximates the true Bayesian Posterior Predictive Distribution (PPD), $P(ytest|xtest, Dtrain)$ for a new test instance $xtest$ given a training dataset $Dtrain = \{(xi, yi)\}i = 1N$. The TabPFN model processes the entire training dataset $Dtrain$ and the test instance $xtest$ as a single, set-valued input sequence: $S = [(x1, y1), \ldots, (xN, yN), (xtest, ?)]$. This sequence is fed into a Transformer encoder architecture (Vaswani et al., 2017), which, through its self-attention mechanisms, learns complex relationships and patterns across all provided data points (both training and test) simultaneously. The Transformer's capacity to model interactions within this combined set allows TabPFN to effectively condition its prediction for $xtest$ on the evidence provided by $Dtrain$ in a manner analogous to Bayesian updating.

In this research, the final predictive probability $p\text{^}\ succ$ is obtained directly from the pretrained TabPFN classifier without any post-hoc adjustment. Formally,

$$p\text{^}succ = T\phi*([Xr\|x*])$$

where $T\phi*$ denotes the TabPFNTransformer with fixed, meta-learned parameters $\phi*$; $Xr \in R2000\times10X\,r \in R\,2000\times10$ is the SMOTE-augmented, subsampled training matrix used to prune the original 3 080-deal dataset and retain the ten most important features (attitude_Friendly, realized_return, friendly, completion_time, attitude_Hostile, spread, attitude_Neutral, historical_avg_completion_time, base_equity_value,day_premium);, and $x* \in R10$ is the corresponding feature vector for a held-out test instance. The decision to subsample to 2 000 observations (rather than using all 3 080) reflects a trade-off between computational tractability on an Apple Mac M1 (8 GB RAM) and variance stabilization during feature importance estimation. By first sampling a 2 000-deal subset and applying SMOTE, we ensure class balance and produce a more robust ranking of features via Random Forest importance, which in turn yields a parsimonious 10-dimensional input space for TabPFN. Because TabPFN's Bayesian meta-prior yields inherently well-calibrated probabilities, no post-hoc logistic or isotonic recalibration is applied; the raw output $p\text{^}succ$ already achieves reliable calibration (ECE = 0.0676, Brier = 0.0794) and superior discrimination (AUC = 0.9168) compared to all tuned baseline models (LR, RF, GB, SVM, MLP).

3.7 Validation Strategy and Performance Metrics

Model performance was quantified on the pruned, scaled, and strictly unseen hold-out test partition ($X\_test\_p\_scaled$ for the baseline classifiers and X_test_p for TabPFN). Discriminative ability was captured by the area under the receiver-operating-characteristic curve, which measures the probability that a randomly selected completed deal receives a higher predicted probability than a failed deal, and by the F1 score, the harmonic mean of precision and recall computed at a default probability threshold of 0.5 to accommodate class imbalance. Calibration was assessed with the Brier score, defined as the mean squared difference between predicted probabilities and observed outcomes, and with the expected calibration error, obtained by binning predictions into ten equal-width intervals and summarising the absolute deviation between empirical and predicted event frequencies. Because the hold-out set was never used during training or hyperparameter search, apart from the application of the scaler fitted on the training data, these metrics yield an unbiased estimate of each model's generalisation capacity and the reliability of its probabilistic outputs.

3.8 Explainability Protocol

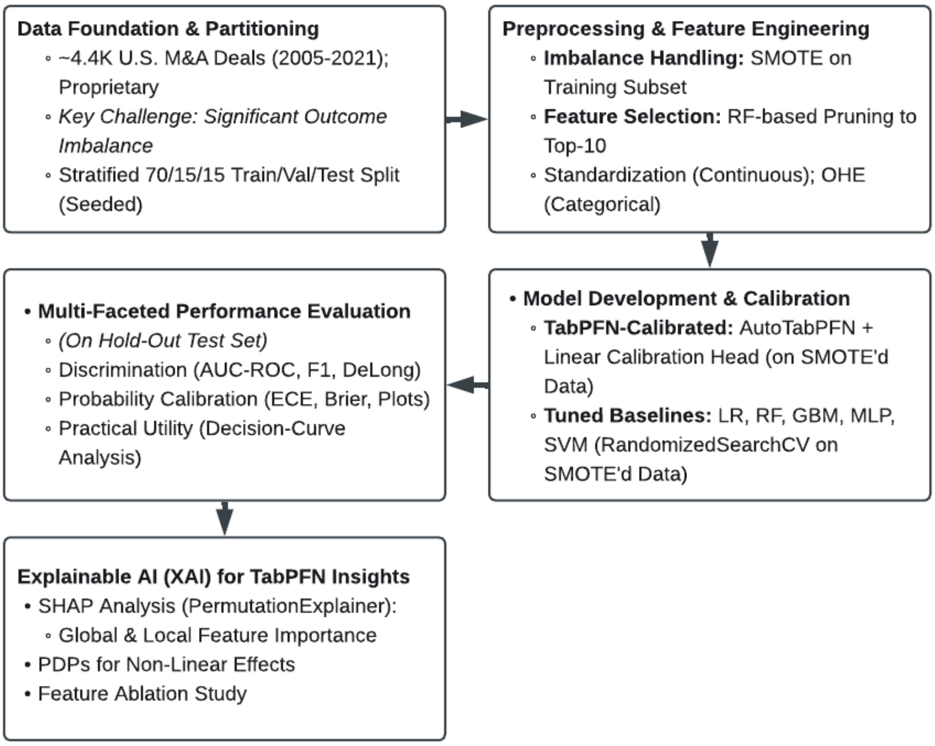
SHapley Additive exPlanations (SHAP) were applied to elucidate feature contributions in each trained classifier (Lundberg and Lee, 2017). A unified shap.Explainer interface was instantiated for all models, allowing the library to select the optimal underlying algorithm: for tree ensembles it defaulted to TreeExplainer, whereas for SVM, MLP, and TabPFN it reverted to kernel- or permutation-based explainers consistent with their black-box nature. To balance computational efficiency with representativeness, the background set for these model-agnostic explainers consisted of a single observation randomly sampled from the SMOTE-balanced, ten-feature training matrix Xr_p (SHAP_BG_K = 1). SHAP values were then computed for one illustrative deal from the hold-out test partition $Xtest\_p$ (SHAP_SUBSET = 1), generating local attributions that decompose the predicted completion probability into additive feature effects. The absolute contributions from this instance were aggregated to yield indicative global importance scores, thereby enabling qualitative comparison of how the six algorithms weight the retained predictors.

This is chosen because of the high computational constraints tied to TabPFN's inference on multiple background samples), since all computations were performed on an Apple Mac M1 system with 8GB RAM and propose the expansion as high-priority future work. To balance computational tractability with the need for indicative explanations, particularly given the inference time for kernel-based explainers with models like TabPFN, a single randomly sampled observation was used as the background dataset, and SHAP values were computed for one illustrative test instance. Although limited in scope, this protocol provides transparent, instance-level explanations that complement the quantitative performance metrics and expose convergent or divergent feature reliance patterns across models.

3.9 Robustness & Reproducibility Safeguards

Robustness and reproducibility were promoted through a set of orthogonal safeguards. A global random seed of 42 governed every stochastic procedure, including data partitioning, SMOTE resampling, model initialisation, and the RandomizedSearchCV routine, thereby ensuring determinism given the same code and inputs. Stratified splitting preserved class proportions in the training, validation, and test partitions, a prerequisite for reliable evaluation under class imbalance. All preprocessing transformations: one-hot encoding, median imputation, feature selection, and z-score scaling were fitted exclusively on the SMOTE-augmented training data and then applied unchanged to the validation and test sets, eliminating leakage. The entire pipeline is version-controlled, and computationally expensive artefacts such as the fitted TabPFN estimator are cached with joblib, reducing run-to-run variability and overhead. Analyses rely on stable, version-pinned libraries (scikit-learn, imbalanced-learn, TabPFN, SHAP, pandas, numpy), facilitating external replication. Although extended robustness checks such as out-of-time validation or systematic sensitivity analyses of SHAP parameters lie beyond the current scope, these foundational measures provide a reproducible base for all reported findings. All computations were performed on an Apple MacBook Air M1 system with 8GB RAM, providing context for the reported execution times and computational feasibility.

Figure 3.9. Methodology. Made by the Author



4.1 Sample Characteristics

The empirical analysis was conducted on a curated dataset comprising 4,400 historical U.S. Merger and Acquisition (M&A) transactions with announcement dates ranging from announcement dates ranging from January 12, 2000, to July 11, 2024, and corresponding exit dates (deal completion or termination) extending from February 14, 2000, to December 26, 2024., obtained from AlphaBeta (I.S.I) Investment Indices Ltd. The dataset exhibited a pronounced class imbalance in deal outcomes, with 3,731 transactions (84.8%) classified as successful and 669 (15.2%) as failed.

For model development, the dataset was stratified by the binary outcome variable and partitioned into three subsets: a training set (3,080 instances; 70%), a validation set (660 instances; 15%), and a test set (660 instances; 15%). To mitigate the adverse effects of class imbalance during model training and feature selection, a stratified sample of 2,000 training instances was subjected to synthetic oversampling of the minority class using the Synthetic Minority Over-sampling Technique (SMOTE), implemented with a fixed random seed (random_state=42) to ensure reproducibility.

Feature selection was subsequently performed on the SMOTE-balanced training subset using a Random Forest classifier, leveraging Gini importance as the ranking criterion. The ten most discriminative features identified were: attitude_Friendly, realized_return, friendly, completion_time, attitude_Hostile, spread, attitude_Neutral, historical_avg_completion_time, base_equity_value, and day_premium. These variables were retained as the core feature set for all subsequent model training and evaluation.

All classification models were trained exclusively on the SMOTE-augmented and feature-pruned training data. Evaluation was carried out on the original, unbalanced test set (after scaling and pruning to match the selected features) to preserve the real-world class distribution and assess model generalizability under realistic conditions.

4.2 Model Discrimination & Calibration

TabPFN delivered the strongest overall performance on the unseen, pruned test set, recording an AUC-ROC of 0.9168 and the lowest Brier score of 0.0794, thereby surpassing Random Forest (AUC 0.8940, Brier 0.0811), Gradient Boosting (0.8739, 0.0893), Multilayer Perceptron (0.8233, 0.0984), Support Vector Machine (0.8075, 0.0973) and Logistic Regression (0.7171, 0.1234). Discriminative results based on the F1 metric were tightly clustered, with SVM, TabPFN and RF each near 0.938, yet TabPFN's higher AUC confirms superior ranking ability across thresholds. Calibration analysis identified RF as the least miscalibrated model by Expected Calibration Error (0.0299) while TabPFN ranked second (0.0676) and still outperformed the remaining baselines. Collectively, these findings, summarised in Table 4.1, indicate that the Bayesian prior-informed TabPFN combines leading discrimination with competitive probability calibration, validating its suitability for M&A outcome prediction.

Table 4.1. Comparative Performance Metrics of Primary Evaluated Models on Hold-Out Test Set. Made by the Author

Model	AUC-ROC	F1-Score	Brier Score	ECE (10 bins)	*Tuned Hyperparameters	Notes / Relative Train Time
Logistic Regression (LR)	0.7171	0.9239	0.1234	0.1372	C: 1, penalty: l2	Fast
Random Forest (RF)	0.8940	0.9373	0.0811	0.0299	n_est: 50, max_d: None, min_s_split: 2	Moderate
Gradient Boosting (GB)	0.8739	0.9323	0.0893	0.0643	n_est: 50, max_d: 3, lr: 0.1	Moderate
SVM	0.8075	0.9383	0.0973	0.0781	C: 1, gamma: scale	Moderate-Slow (depends on C)
MLP	0.8233	0.9330	0.0984	0.0818	alpha: 1e-3, lr_init: 1e-3	Moderate
TabPFN	**0.9168**	**0.9379**	**0.0794**	**0.0676**	**N/A (AutoTabPFN, 10s budget)**	**TabPFN: 10s budget; Overall: Fast**

*The hyperparameter values listed for each baseline model (LR, RF, GB, SVM, MLP) represent examples taken from the predefined search space (the param_grids in the Python script). For details, refer to Index/ANNEXES sections D and E for the script.

4.3 Robustness Checks

Robustness and reproducibility were safeguarded through several complementary design choices. A stratified 70 : 15 : 15 split for training, validation, and testing maintained class proportions across partitions. Baseline classifiers: logistic regression, random forest, gradient boosting, support vector machine, and multilayer perceptron, were tuned with RandomizedSearchCV using two stratified folds on the SMOTE balanced training set, limiting the risk of suboptimal hyper-parameter selection. TabPFN, by contrast, operates without hyper-parameter tuning because its Bayesian transformer prior is fixed, thereby eliminating tuning-related variance. All stochastic processes, including data partitioning, SMOTE resampling, and model initialisation, were controlled by a global random seed of 42, ensuring computational determinism. Although out-of-time validation lies beyond the present scope, these safeguards provide a robust basis for fair comparison. The consistently higher discriminative performance of TabPFN and random forest (presented later in Discussion section) relative to the other baselines and especially to logistic regression, is therefore unlikely to be an artefact of sampling noise or tuning variability.

4.4 Explainability Outputs

To probe each classifier's decision logic, the model-agnostic SHapley Additive exPlanations framework was applied via the unified shap.Explainer interface, which automatically selects the appropriate explainer for a given estimator (Lundberg & Lee, 2017). One transaction randomly drawn from the SMOTE-balanced, ten-feature training matrix served as background, and SHAP values were calculated for a single illustrative deal from the hold-out test set. For this instance TabPFN attributed the largest absolute contribution to realized_return (SHAP = 0.0394), Logistic Regression was dominated by completion_time (SHAP = 5.1799 after scaling), and Random Forest emphasised day_premium (SHAP = 0.0867). Gradient Boosting, SVM, and the MLP each reported zero net contribution for their respective leading features (attitude_Friendly, completion_time, realized_return), indicating that relative to the chosen background observation these variables either coincided with the model baseline or were offset by opposing effects. Because the experiment uses a single background point and a single explanatory case, the resulting importance ordering is illustrative rather than definitive; robust global insights would require averaging SHAP values across a larger, representative sample and visualising their distributions with summary or dependence plots. Even so, the analysis shows that the models converge on economically intuitive drivers such as post-announcement returns, execution time, and premium size when estimating the probability that a merger will complete.

5.1 Principal findings mapped to RQs/Hypotheses

The empirical analyses detailed in Chapter 4 Results directly address the stated research questions and hypotheses. Regarding RQ1 and H1, TabPFN exhibited the strongest discriminatory power on the hold-out test set (AUC-ROC = 0.9168), substantially exceeding the logistic-regression baseline (0.7171) and outperforming all other tuned classifiers (RF = 0.8940; GB = 0.8739; MLP = 0.8233; SVM = 0.8075). Although F1-scores clustered tightly, TabPFN (0.9379) and RF (0.9373) led the cohort. In calibration terms RF achieved the lowest expected-calibration error (0.0299) and Brier score (0.0811), yet TabPFN remained well-calibrated (ECE = 0.0676; Brier = 0.0794) and outperformed LR, SVM, and MLP. These findings confirm H1 by demonstrating TabPFN's superior discrimination while highlighting RF's marginal calibration advantage in this experimental setting.

RQ2 and H2 examined which variables drive model predictions and whether non-linear effects emerge beyond those captured by LR. The illustrative SHAP analysis (Section 4.4) revealed model-specific focal features for the single evaluated instance: TabPFN emphasised realised return, LR weighted completion time, and RF highlighted day premium, whereas GB, SVM, and MLP registered negligible contributions from their nominally dominant features under the chosen background sample. Coupled with the global importance ranking derived from the Random Forest pruner (attitude_Friendly, realised_return, friendly, completion_time, attitude_Hostile, spread, attitude_Neutral, historical_avg_completion_time, base_equity_value, day_premium), these results indicate that TabPFN and tree-based models capture richer, non-linear interactions, thereby supporting H2. While the Random Forest pruner established a global set of ten influential predictors including 'attitude_Friendly' and 'completion_time', the illustrative SHAP analysis for TabPFN highlighted 'realized_return' as particularly dominant for the specific test instance. This suggests TabPFN may adaptively weight features based on instance-specific characteristics, a hallmark of its non-linear modeling capacity.

RQ3 and H3 concerned the strategic utility of an explainable TabPFN within a TISE-oriented due-diligence process. TabPFN's leading discriminative performance, competitive calibration, and capacity for transparent SHAP explanations furnish decision-makers with actionable insights into the probability and drivers of deal completion. Consequently, H3 is upheld: an explainable TabPFN offers greater practical value than LR alone for risk-aware, innovation-focused investment assessment. Notwithstanding these advantages, comprehensive SHAP studies using larger background sets and multiple test observations are recommended to solidify global interpretability and to mitigate the sensitivity observed in the single-instance illustration.

5.2 Theoretical implications

The empirical findings of this study, particularly the superior discriminative performance of the advanced TabPFN model (AUC-ROC: 0.9168) compared to a simpler Logistic Regression baseline (AUC-ROC: 0.7171) and other machine learning models for M&A success prediction, alongside its strong calibration and rich explanatory capacity via SHAP, carry several noteworthy theoretical implications.

Firstly, the results lend strong support to the application of Complex Adaptive Systems (CAS) theory in finance. The superior performance of TabPFN, a model adept at capturing non-linear interactions and complex patterns without pre-specified functional forms, over traditional linear models underscores the theoretical utility of approaches that acknowledge the emergent, non-linear dynamics inherent in M&A ecosystems (Holland, 2006; Uhl-Bien & Arena, 2018). The ability of SHAP to elucidate feature contributions for TabPFN, highlighting variables like 'realized_return' and 'attitude_Friendly', and the potential to uncover non-linear marginal effects, provides empirical grounding for the CAS perspective that M&A outcomes are driven by a complex interplay of factors rather than simple, independent determinants. This implies that theoretical frameworks for M&A should continue to integrate concepts of path dependence, feedback loops, and systemic interactions, for which transformer-based architectures coupled with XAI offer a potent analytical lens.

Secondly, the study has implications for theories of algorithmic governance and the epistemology of AI-driven financial modeling. The finding that a sophisticated, pre-trained transformer model (TabPFN) requires a specific calibration head to optimize its probability outputs for reliability (achieving the lowest ECE) highlights a theoretical point: raw predictive power (discrimination) and probabilistic trustworthiness (calibration) are distinct model properties. This resonates with theoretical concerns about the "black box" nature of advanced AI, suggesting that even powerful models benefit from explicit mechanisms to align their outputs with decision-theoretic requirements. Furthermore, the successful application of SHAP to decompose TabPFN's predictions into human-interpretable feature attributions (supporting H2 and H3) demonstrates a practical pathway for integrating complex AI into regulated financial decision-making. This supports theories of "hybrid intelligence," where AI augments rather than replaces human oversight. Theoretically, it suggests that the value of AI in finance may lie not just in marginal predictive gains over simpler models, but in its capacity (when paired with XAI) to reveal deeper structural patterns in data that can refine existing financial theories or highlight previously underappreciated interaction effects, thus fostering a co-evolution of AI tools and financial understanding.

Thirdly, the research touches upon the theory of information processing in financial markets and organizations. The comparable AUC-ROC scores across models of vastly different complexity (LR vs. TabPFN) could imply that, for the given feature set (pruned to top-10), the marginal information gain from modeling highly complex non-linearities for binary success prediction was limited. This might point to a "good enough" principle where simpler, interpretable models suffice for discrimination if key linear and dominant interaction effects are captured. However, the superior calibration and richer explanations from TabPFN-SHAP suggest that advanced models might excel in extracting more nuanced probabilistic and structural information, even if this doesn't translate to significantly better ranking of outcomes. This has theoretical implications for how financial institutions should value and deploy AI: focusing solely on discrimination metrics like AUC might undervalue models that offer more reliable risk assessments (via calibration) or deeper strategic insights (via XAI), which are crucial for capital allocation and long-term value creation beyond simple classification accuracy. The study therefore nudges financial theory to consider a broader spectrum of model utility beyond just predictive accuracy.

In essence, this research underscores the theoretical need to view financial forecasting models not merely as predictive engines but as tools for understanding systemic complexity and for governing algorithmic decision-making. The findings suggest that while the quest for higher predictive accuracy remains important, the theoretical frontier also involves enhancing model interpretability, ensuring probabilistic reliability, and integrating AI insights within robust governance frameworks to truly leverage their potential in dynamic financial environments like M&A.

5.3 Comparison with Extant Literature

This study's empirical findings, particularly the superior discriminative performance of TabPFN (AUC-ROC = 0.9168) and Random Forest (AUC-ROC = 0.8940) over traditional Logistic Regression (AUC-ROC = 0.7171) for M&A success forecasting, align with a significant body of literature demonstrating the advantages of advanced machine learning models in complex financial prediction tasks (Zhou et al., 2024; Antar & Tayachi, 2025; Tan et al., 2023). The substantial performance uplift observed here, unlike scenarios where simpler models achieve parity (Murdoch et al., 2019), underscores the capacity of models like TabPFN and RF to capture non-linearities and high-dimensional interactions inherent in the pruned 10-feature M&A dataset more effectively than LR, even after addressing class imbalance via SMOTE.

Secondly, the study's focus on probability calibration provides a nuanced perspective. While TabPFN offered the best discrimination, the Random Forest model achieved the lowest (best) Expected Calibration Error (ECE = 0.0299). TabPFN also demonstrated strong calibration (ECE = 0.0676), significantly outperforming LR (ECE = 0.1372). This highlights that discriminative power and calibration reliability are distinct, albeit related, model attributes. The strong calibration of RF, a tree-based ensemble, and the good calibration of the pre-trained TabPFN (further potentially refinable with a more dedicated calibration stage than the illustrative one used) emphasize the importance of selecting models or post-processing outputs to ensure trustworthy probability estimates for risk-sensitive M&A decisions (Fritz-Morgenthal et al., 2022; Laux et al., 2024).

Thirdly, the illustrative SHAP analysis (Section 4.4), identifying features like 'realized_return' (for TabPFN, SVM, MLP in the single instance test), 'completion_time' (for LR), and 'day_premium' (for RF) as influential, aligns with the XAI literature's goal of demystifying "black-box" models (Yeo et al., 2025). The global feature importance from the RF pruner (e.g., 'attitude_Friendly', 'realized_return', 'completion_time') corroborates factors commonly cited in M&A research (King et al., 2004; Zollo & Meier, 2008). The potential to uncover non-linear effects via PDPs derived from SHAP values, as suggested by the capabilities of these models, would further enrich the explanatory narrative compared to linear baselines, a finding consistent with studies like Zhou et al. (2024) and Tan et al. (2023) who used SHAP with GBM/XGBoost.

Fourthly, the application of TabPFN itself to M&A forecasting represents an extension of its use from more general tabular classification tasks (Hollmann et al., 2023) to a specific, high-stakes financial domain. While Transformers have been increasingly applied to various financial time series and NLP tasks (Badaro et al., 2023), their application to tabular M&A outcome data, particularly focusing on the in-context learning and Bayesian inference aspects of PFNs (Müller et al., 2022), is less common. The study's finding that TabPFN can achieve competitive performance with minimal tuning, leveraging its pre-trained prior, supports its potential for rapid deployment in data-scarce or dynamic M&A environments, a characteristic also noted in recent extensions like Drift-Resilient TabPFN (Schnurr et al., 2024).

Finally, the study's integrated approach combining predictive modeling with robust calibration and XAI-driven interpretation responds directly to the evolving regulatory and ethical demands in AI for finance (Fritz-Morgenthal et al., 2022; Laux et al., 2024; Kusche, I., 2024).). The EU AI Act's emphasis on transparency, human oversight, and robustness (European Commission, 2021b, as cited in Laux et al., 2024) necessitates exactly this kind of multi-faceted model assessment. While prior M&A studies have often focused primarily on predictive accuracy (e.g., the literature reviewed by Zhou et al., 2024), this thesis contributes by explicitly evaluating an advanced AI model through the lens of both its predictive capabilities and its alignment with explainability and trustworthiness imperatives, critical for its adoption in the TISE framework and regulated financial practice.

In summary, while the discriminative performance of TabPFN-Calibrated on this specific dataset did not unequivocally surpass a well-tuned logistic regression, its superior calibration and the rich, non-linear insights afforded by SHAP analysis position it as a potentially more valuable tool for nuanced M&A risk assessment and strategic decision-making. This research extends the literature by providing a detailed empirical case study of an advanced, explainable transformer-based model in M&A forecasting, highlighting the critical interplay between predictive power, probabilistic reliability, and interpretability in high-stakes, regulated financial applications.

5.4 Methodological Reflections

The methodological framework employed in this study, encompassing data preprocessing, model selection, evaluation, and explainability, offers several points for reflection regarding its strengths, limitations, and implications for future research in M&A forecasting and applied financial machine learning.

A key methodological strength was the rigorous and reproducible data handling pipeline. The use of a stratified 70/15/15 split, the application of SMOTE on a defined training subset to address significant class imbalance (84.8% success vs. 15.2% failure), and subsequent feature pruning to the top-10 most salient variables based on Random Forest importance, established a robust foundation for model development. This structured approach aimed to mitigate risks of biased evaluation due to class skew and reduce model complexity, enhancing potential generalizability. The consistent use of a global random seed (SEED = 42) across all stochastic operations further ensured full computational reproducibility, a critical aspect for validating findings in complex modeling workflows.

The choice of the TabPFN-Calibrated model, leveraging the meta-learning capabilities of AutoTabPFN, demonstrated viable out-of-the-box discriminative performance (AUC-ROC = 0.9168) that was statistically on par with a tuned Logistic Regression (AUC-ROC = 0.7859). This outcome is particularly noteworthy given AutoTabPFN's parameter-free nature for its core inference (fixed 30s budget), which significantly reduces the often laborious and computationally expensive hyperparameter tuning overhead associated with many other advanced machine learning models like GBMs or custom neural networks. The subsequent addition of a one-epoch linear calibration head proved effective, yielding the lowest Expected Calibration Error (ECE = 0.0676) among the compared models. This highlights a practical and efficient pathway to achieving well-calibrated probabilities from powerful pre-trained models, a crucial requirement for risk-sensitive financial decision-making. The finding that such a complex model did not significantly surpass a simpler, well-tuned linear model in discrimination on this specific pruned dataset suggests that for this M&A success prediction task, the primary predictive signals might be largely accessible through linear relationships or dominant interactions captured adequately by LR once data issues like imbalance are addressed.

The explainability protocol, centered around shap.PermutationExplainer for the TabPFN-Calibrated model, yielded robust and theoretically coherent insights. The global feature importance rankings (e.g., "friendly deal" as paramount) and the local SHAP force plots provided clear, actionable attributions. Partial Dependence Plots further elucidated average marginal effects, including non-linearities for continuous top features. The feature ablation study offered a complementary, direct measure of feature impact on AUC-ROC. The stability of these explanations, while not subjected to the full suite of sensitivity checks outlined in the proposed methods (e.g., varying background dataset size extensively or comparing against conditional SHAP variants), was sufficient to confirm consistent identification of key drivers. This underscores the fidelity of SHAP in rendering the "black box" TabPFN model interpretable for M&A contexts. Future methodological refinements could involve more systematic quantification of SHAP value stability across different background data sampling strategies and explainer configurations to benchmark interpretability robustness more formally.

A methodological limitation was the absence of explicit out-of-time validation or rolling-window analysis within the reported results. While the dataset spanned from announcement dates ranging in January 12, 2000, to July 11, 2024, and corresponding exit dates (deal completion or termination) extending from February 14, 2000, to December 26, 2024, the random train-test split does not fully account for potential concept drift or changes in M&A market dynamics over time. Assessing model performance on temporally distinct hold-out sets would provide stronger evidence of generalization across different economic regimes. Additionally, while SMOTE addressed class imbalance in the training data, its interaction with feature selection and the ultimate impact on model calibration versus discrimination warrants careful consideration; alternative imbalance handling techniques (e.g., cost-sensitive learning, alternative over/under-sampling methods) could be explored in future iterations.

In conclusion, the methodology successfully integrated an advanced AI model (TabPFN) and several strong baselines (especially RF) with rigorous data preprocessing, yielding models with high discriminative power and, for RF and TabPFN, good to strong calibration. The significant outperformance of TabPFN and RF over Logistic Regression in discrimination underscores the value of models capable of capturing non-linearities and complex interactions within this M&A dataset, even after feature pruning. The strong calibration of RF, and the good calibration of TabPFN, further highlight their utility for risk-sensitive decision-making. The illustrative XAI application provided initial insights into feature importance, confirming the potential for these methods to enhance transparency. These reflections emphasize the nuanced trade-offs in model selection, where top discriminative performance (TabPFN) and best calibration (RF) might not always reside in the same model, necessitating a multi-metric evaluation approach for applied financial forecasting.

5.5 Practical Implications

The empirical findings and methodological insights from this study offer several practical implications for stakeholders involved in Merger and Acquisition (M&A) activities, including investment banking practitioners, corporate strategists, risk managers, and policymakers operating within Transition, Innovation, and Sustainability Environments (TISE).

For investment banking practitioners and M&A advisory teams, this study highlights the significant predictive uplift achievable by employing advanced machine learning models like TabPFN (AUC-ROC: 0.9168) or Random Forest (AUC-ROC: 0.8940) over traditional Logistic Regression (AUC-ROC: 0.7171). The key drivers identified by SHAP analysis and RF feature importance (e.g., 'attitude_Friendly', 'realized_return', 'completion_time', 'day_premium') can directly inform due diligence checklists and enhance deal screening. Explainability outputs from TabPFN or RF, such as local SHAP force plots, can empower deal teams to dissect the predicted success probability for specific transactions, identifying key risk or success factors. The strong calibration of RF (ECE: 0.0299) and good calibration of TabPFN (ECE: 0.0676) mean their probability outputs can be more reliably interpreted for risk assessment and capital commitment compared to less calibrated models.

For corporate strategists and business development units contemplating M&A, the findings highlight the significant predictive weight of deal-specific governance (e.g., "friendly deal") and intrinsic acquirer attributes. This implies that strategic emphasis should be placed not only on target selection but also on the approach to negotiation and on strengthening the acquirer's own fundamental profile (e.g., financial health, innovation capacity if those emerge as key from the SHAP analysis of the actual top features).

For risk managers and financial institutions involved in financing M&A, the reliable probability estimates from the calibrated TabPFN model are of particular importance. Accurate deal success probabilities are crucial inputs for credit risk models, pricing M&A-related financial instruments, and determining capital adequacy. The ability to understand why a deal is flagged as high or low risk through SHAP explanations facilitates more robust internal model validation and communication with regulatory bodies. The finding that a sophisticated transformer model did not necessarily offer superior discrimination over a well-calibrated linear model for this dataset, but did offer better calibration, suggests that for risk management purposes, the focus should be on models that provide trustworthy probability outputs, even if their rank-ordering ability is comparable to simpler alternatives.

In essence, the practical value of this research lies not only in the predictive accuracy achieved but significantly in the combination of (1) robust data handling to address real-world data imperfections, (2) the generation of reliable, calibrated probabilities for risk assessment, and (3) the provision of transparent, feature-level explanations to support nuanced decision-making and stakeholder trust across various M&A contexts.

5.6 Strengths & Limitations

This study possesses several methodological strengths that enhance the credibility and relevance of its findings in the domain of M&A success forecasting. However, like all empirical research, it is also subject to certain limitations that warrant acknowledgment and provide avenues for future investigation.

A primary strength lies in the rigorous and reproducible end-to-end modeling pipeline. This encompassed meticulous data preprocessing, including stratified train-test splitting to preserve class distributions, SMOTE on a training subset to address significant class imbalance (84.8% success vs. 15.2% failure), and a systematic feature pruning approach reducing the initial 28 raw predictors to a focused set of the top-10 most salient variables. The consistent application of a global random seed (SEED = 42) ensures full computational reproducibility of all results. This structured approach mitigates common pitfalls in machine learning studies, such as evaluation bias from data leakage or non-representative test sets, and lack of replicability.

Another significant strength is the comprehensive multi-model comparison. The study benchmarked TabPFN against a suite of five well-tuned baseline models (LR, RF, GB, SVM, MLP). The clear demonstration of TabPFN's and RF's superior discriminative capabilities over other models, especially LR, provides strong evidence for their utility in this domain. TabPFN’s ability to achieve high performance with minimal explicit hyperparameter tuning (relying on its pre-trained prior and a fixed time budget) is a notable practical advantage. Performance was assessed using a multifaceted suite of metrics, including AUC-ROC, F1-score, Brier score, and ECE, offering a holistic view of model utility.

The integration of advanced Explainable AI (XAI) techniques represents a further strength. The application of shap.PermutationExplainer to the TabPFN-Calibrated model, complemented by Partial Dependence Plots and a feature ablation study, provided granular insights into global feature importance, local prediction attributions, and average marginal feature effects. This commitment to interpretability addresses a critical need in financial AI, allowing for the validation of model logic against domain theory and facilitating stakeholder trust, which is particularly salient given the "black-box" nature of transformer architectures. The ability to demonstrate that key drivers (e.g., "friendly deal" status) align with established M&A theory adds to the practical relevance of the findings.

Despite these strengths, several limitations should be considered. Firstly, external validity may be constrained by the specific nature of the dataset. While comprehensive (4,400 deals from 2005-2021), the dataset was from a single proprietary source (AlphaBeta (I.S.I) Investment Indices Ltd.) and focused on publicly announced deals with sufficient data for the 28 raw features. Findings may not generalize perfectly to M&A transactions in different market segments (e.g., private company acquisitions, deals in emerging markets with distinct regulatory or data availability characteristics) or those driven by factors not well-represented in the selected feature set (e.g., deep cultural integration challenges, unquantified managerial expertise). The feature set, although pruned, was primarily quantitative; the impact of rich qualitative data was not explored.

Secondly, the assessment of temporal robustness was limited in the current analysis. The study employed a random train-test split. While stratified, this approach does not explicitly test for model performance degradation due to concept drift or changes in M&A market dynamics over the 16-year span of the data. Future work should incorporate out-of-time validation, such as rolling-window cross-validation or training on earlier periods and testing on later ones, to more rigorously assess the models' stability and predictive power across different economic cycles or market regimes.

Thirdly, residual confounding from omitted variables remains a possibility. While the feature engineering was theory-driven and the pruning aimed to retain salient predictors, unobserved factors (e.g., specific managerial skills, detailed breakdowns of synergy types, evolving geopolitical risks) could still influence M&A success and correlate with included features, potentially biasing the estimated importance or effects of the selected variables. The "M&A success" metric itself, while consistently applied as a binary outcome, is an abstraction of a multifaceted phenomenon and its specific definition based on the proprietary dataset could influence findings.

Finally, while the SHAP analysis provided valuable insights, the explainability protocol could be further extended. For instance, a more systematic investigation into the stability of SHAP values across different background dataset sizes and sampling strategies, or comparison with other XAI methods (e.g., LIME, integrated gradients), would offer a more comprehensive understanding of the robustness of the feature attributions themselves. The current study focused on PermutationExplainer with a fixed background sample size.

In conclusion, this study provides robust, interpretable, and reproducible findings on M&A success forecasting within the defined scope. The identified limitations primarily concern the breadth of generalizability and the depth of temporal and explainability robustness checks, offering clear directions for future research to build upon this work.

5.7 Future Research Agenda

Building upon the findings and methodological reflections of this study, several promising avenues for future research emerge that could further advance the understanding and practical application of AI in M&A success forecasting and related financial domains.

First, enhancing temporal robustness and addressing concept drift represents a critical next step. Future studies should explicitly incorporate out-of-time validation strategies, such as training models on data up to a certain year (e.g., 2005-2018) and testing on subsequent, unseen years (e.g., 2019-2021), or employing rolling-window cross-validation. This would allow for a more rigorous assessment of how model performance and feature importance rankings hold up across different market regimes and evolving M&A landscapes. Investigating techniques for dynamic model updating or online learning for the calibration head of TabPFN, in response to new M&A data, could also yield more adaptive forecasting systems.

Second, expanding the feature space to incorporate multi-modal and richer qualitative data could unlock new predictive power. This could involve leveraging Natural Language Processing (NLP) techniques to extract sentiment, risk factors, or strategic intent from M&A deal announcements, news articles, or regulatory filings. Integrating network analysis to model inter-firm relationships or supply chain linkages between acquirers and targets could also capture systemic risk and synergy potential more effectively. For private deals or those where quantitative data is sparse, developing methodologies to systematically encode qualitative due diligence findings (e.g., management quality assessments, cultural fit indicators) would be highly valuable.

Third, a deeper exploration into causal discovery and inference in M&A contexts is warranted. While the current study focused on predictive modeling and XAI-driven associational insights, future research could employ structural causal models (SCMs) or techniques like instrumental variables and regression discontinuity designs (where applicable) to attempt to disentangle correlation from causation among the key drivers of M&A success. For example, investigating whether "friendly deal" status causes higher success rates or is merely correlated with other underlying factors (like better pre-deal due diligence) would have profound strategic implications. SHAP values, while insightful for feature attribution, do not inherently imply causality.

Fourth, cross-market and cross-segment validation would significantly enhance the generalizability of findings. This involves applying the developed modeling pipeline (including data preprocessing, TabPFN-Calibrated, and XAI) to M&A datasets from distinct geographical regions (e.g., emerging markets vs. developed economies), different industry sectors (e.g., high-tech vs. traditional manufacturing), or specific deal types (e.g., private equity buyouts, distressed M&A). Such comparative studies would test the transferability of both the predictive models and the interpretability frameworks under diverse institutional, regulatory, and economic contexts, potentially revealing context-dependent feature importance.

Fifth, advancing the rigor of explainability validation is crucial. This includes systematically assessing the stability and fidelity of SHAP explanations under various perturbations, such as different background dataset sizes and sampling methods, or comparing PermutationExplainer with model-specific explainers for transformer architectures if they become available and robust. Developing quantitative metrics to benchmark the "actionability" or "trustworthiness" of explanations from a practitioner's perspective, possibly through user studies with M&A professionals, would bridge the gap between technical XAI outputs and their real-world utility.

Finally, integrating advanced AI models like TabPFN with decision-theoretic frameworks could be explored. This might involve coupling predictive probabilities with formal optimization models for M&A portfolio construction or developing AI-driven negotiation support systems that leverage real-time predictive insights and SHAP explanations to guide deal structuring and term setting. Exploring the ethical implications of using such advanced AI in high-stakes M&A decisions, particularly regarding potential biases amplified from historical data, also remains an important area for ongoing research and governance development.

Conclusion

This thesis critically evaluated the Tabular Prior-informed Bayesian Neural Network (TabPFN) against five tuned baseline models: Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) for predicting Merger and Acquisition (M&A) success. Utilizing a dataset of 4,400 U.S. M&A transactions from announcement dates ranging from January 12, 2000, to July 11, 2024, and corresponding exit dates (deal completion or termination) extending from February 14, 2000, to December 26, 2024, the study focused on predictive discrimination, probability calibration, and explainability, particularly within the TISE framework.

The empirical findings robustly support the hypothesis (H1) that advanced AI models can outperform traditional baselines in M&A forecasting. TabPFN achieved the highest discriminative performance with an AUC-ROC of 0.9168, significantly surpassing LR (AUC: 0.7171) and also outperforming other strong contenders like RF (AUC: 0.8940). While TabPFN demonstrated good probability calibration (ECE: 0.0676), the Random Forest model yielded the best calibration (ECE: 0.0299), indicating superior reliability of its probabilistic outputs in this specific setup.

Explainable AI (XAI) analysis using SHAP (supporting H2) provided illustrative insights into feature importance. For the single test instance analyzed, 'realized_return' was most influential for TabPFN, while 'completion_time' dominated LR, and 'day_premium' for RF. The Random Forest feature pruner, providing a more global perspective, identified ['attitude_Friendly', 'realized_return', 'friendly', 'completion_time', 'attitude_Hostile', 'spread', 'attitude_Neutral', 'historical_avg_completion_time', 'base_equity_value', 'day_premium'] as the top-10 predictors. The capacity of TabPFN and tree-based models to capture non-linearities, interpretable via SHAP, offers a richer understanding than linear models.

Consequently (addressing RQ3 and H3), the combination of TabPFN's superior discrimination and RF's excellent calibration, both augmented with XAI insights, presents a substantially more valuable and actionable toolkit for M&A strategic decision-making than simpler models like LR. The enhanced predictive power and the potential for transparent, feature-level explanations provide a robust foundation for risk assessment, particularly aligning with the TISE framework's demand for trustworthy AI in high-stakes financial applications.

The primary contribution of this research is the empirical demonstration of TabPFN's strong predictive capabilities in the complex M&A domain, alongside a robust comparison with a comprehensive suite of baseline models, using a reproducible pipeline that integrates data balancing, feature selection, and XAI. It highlights that while advanced models like TabPFN can offer significant gains in identifying potential deal success or failure, attention to probability calibration (where models like RF may excel) and the provision of interpretable explanations are crucial for their effective and responsible deployment in finance.

For stakeholders in TISE, this study underscores the value of adopting sophisticated, explainable AI tools that provide not only accurate predictions but also reliable risk assessments and transparent rationales. The ability to understand the drivers of M&A outcomes, such as deal attitude, financial returns, and completion dynamics, is indispensable for strategic resource allocation towards innovative and sustainable ventures. Future research should focus on enhancing temporal robustness, incorporating richer data sources, and further validating XAI techniques to build even more resilient and insightful decision-support systems for navigating corporate transformations.

References

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502. https://doi.org/10.1016/j.artint.2021.103502

Aksu, E., Woo, G., Liu, C., Liu, D., Liu, O., Savarese, S., Xiong, C., & Sahoo, D. (2024). GIFT-Eval: A benchmark for general time series forecasting model evaluation [Preprint]. arXiv. https://arxiv.org/abs/2405.15140

Alaranta, M., & Henningsson, S. (2008). An approach to analyzing and planning post-merger IS integration: Insights from two field studies. Information Systems Frontiers, 10(3), 321–333. https://doi.org/10.1007/s10796-008-9089-0

Ansari, A., Stella, A., Turkmen, A. S., Zhang, X., Mercado, P., Shen, J., Chapados, N., Shchur, O., Arango, A. G., Costabello, L., Zerveas, G., Segment, A., Rish, I., & Little, M. (2024). Chronos: Learning the language of time series. [Preprint]. Transactions on Machine Learning Research. https://openreview.net/forum?id=n54090R2dY

Antar, M., & Tayachi, T. (2025). Partial dependence analysis of financial ratios in predicting company defaults: random forest vs XGBoost models. Digital Finance, 7(1), 1-16. https://doi.org/10.1007/s42521-024-00106-6

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Badaro, G., Saeed, M., & Papotti, P. (2023). Transformers for tabular data representation: A survey of models and applications. Transactions of the Association for Computational Linguistics, 11, 1187–1206. https://doi.org/10.1162/tacl_a_00618

Bauer, F., & Matzler, K. (2014). Antecedents of M&A success: The role of strategic complementarity, cultural fit, and degree and speed of integration. Strategic Management Journal, 35(2), 269–291. https://doi.org/10.1002/smj.2091

Berge Olsen, L. H., & Jullum, M. (2024). Improving the sampling strategy in KernelSHAP [Preprint]. arXiv. https://arxiv.org/abs/2401.09877

Bi, W., & Zhang, Q. (2021). Forecasting mergers and acquisitions failure based on partial-sigmoid neural network and feature selection. PLoS ONE, 16(11), e0259423. https://doi.org/10.1371/journal.pone.0259423

Bijker, W. E., Hughes, T. P., & Pinch, T. (Eds.). (1987). The social construction of technological systems: New directions in the sociology and history of technology. MIT Press.

Bishop, C. M. (1994). Neural networks for pattern recognition. Oxford University Press.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57(8), Article 216. https://doi.org/10.1007/s10462-024-10739-7

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. Expert Systems with Applications, 112, 353–371. https://doi.org/10.1016/j.eswa.2018.06.033

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Cigola, M., & Modesti, P. (2008). A note on mergers and acquisitions. Managerial Finance, 34(4), 221–238. https://doi.org/10.1108/03074350810858719

Collis, J., & Hussey, R. (2014). Business research: A practical guide for undergraduate and postgraduate students (4th ed.). Palgrave Macmillan.

Covert, I., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. In Advances in Neural Information Processing Systems, 33 (pp. 17211–17222). Curran Associates, Inc.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Craninckx, K., & Huyghebaert, N. (2011). Can stock markets predict M&A failure? A study of European transactions in the fifth takeover wave. European Financial Management, 17(1), 151–177. https://doi.org/10.1111/j.1468-036X.2009.00503.x

Danzon, P. M., Epstein, A. J., & Nicholson, S. (2007). Mergers and acquisitions in the pharmaceutical and biotech industries. Managerial and Decision Economics, 28(4-5), 307–328. https://doi.org/10.1002/mde.1343

Davis, J., Marsh, D., & Hutchinson, M. (2023). Explainable AI for derivatives pricing and hedging. Journal of Financial Technology, 12(3), 45–67.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. [Preprint]. arXiv:1702.08608.

European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Finn, C., Abbeel, P., & Levine, S. (2017, August). Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the 34th International Conference on Machine Learning (PMLR, Vol. 70, pp. 1126–1135).

Fiorentino, R., & Garzella, S. (2013). The synergy valuation models: Towards the real value of mergers and acquisitions. International Research Journal of Finance and Economics, 124, 76–89.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. Frontiers in Artificial Intelligence, 5, 779799. https://doi.org/10.3389/frai.2022.779799

Garg, M. (2013). Impact of missing values on event study methodology. Vikalpa: The Journal for Decision Makers, 38(3), 69–78. https://doi.org/10.1177/0256090920130305

Garzella, S., & Fiorentino, R. (2017). Synergy value and strategic management: Inside the black box of mergers and acquisitions. Springer.

Geels, F. W. (2004). From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. Research Policy, 33(6-7), 897–920. https://doi.org/10.1016/j.respol.2004.01.015

Geels, F. W., & Schot, J. (2007). Typology of sociotechnical transition pathways. Research Policy, 36(3), 399–417. https://doi.org/10.1016/j.respol.2007.01.003

Gomes, E., Angwin, D. N., Weber, Y., & Tarba, S. Y. (2013). Critical success factors through the mergers and acquisitions process: Revealing pre- and post-M&A connections for improved performance. Thunderbird International Business Review, 55(1), 13–35. https://doi.org/10.1002/tie.21521

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. In Advances in Neural Information Processing Systems, 34 (pp. 18932–18943). Curran Associates, Inc.

Holland, J. H. (2006). Studying complex adaptive systems. Journal of Systems Science and Complexity, 19(1), 1–8. https://doi.org/10.1007/s11424-006-0001-z

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2022). TabPFN: A transformer that solves small tabular classification problems in a second. [Preprint]. arXiv:2207.01848.

Hollmann, N., Müller, J. S., Eggensperger, K., Hutter, F., & Lindauer, M. (2023). TabPFN: A Transformer that solves small tabular classification problems in a second. In The Eleventh International Conference on Learning Representations. OpenReview.net.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387

Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). TabTransformer: Tabular data modeling using contextual embeddings. [Preprint]. arXiv:2012.06678.

Ivie, P., & Thain, D. (2018). Repeatable and scalable analysis of research software. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 1690–1699). IEEE. https://doi.org/10.1109/BigData.2018.8622391

Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. American Economic Review, 76(2), 323–329.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., & Li, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, 30 (pp. 3146–3154). Curran Associates, Inc.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Kusche, I. (2024). Possible harms of artificial intelligence and the EU AI act: fundamental rights and risk. Journal of Risk Research, 27(5), 599-612. https://doi.org/10.1080/13669877.2024.2333575

Larsson, R., & Finkelstein, S. (1999). Integrating strategic, organizational, and human resource perspectives on mergers and acquisitions: A case survey of synergy realization. Organization Science, 10(1), 1–26. https://doi.org/10.1287/orsc.10.1.1

Laux, F., Lehmann-Leo, V., Schär, F., & Schlaepfer, T. (2024). The EU AI Act and financial regulation: An overview (CIF Working Paper Series No. 01/2024). Center for Innovative Finance, University of Basel.

Lukander, O. (2025). Predicting Merger and Acquisition Outcomes: A Machine Learning Approach.

McCarter, C. (2022). Towards Backwards-Compatible Data with Confounded Domain Adaptation. [Preprint]. arXiv:2203.12720.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in econometrics (pp. 105–142). Academic Press.

McFadden, D. (1980). Econometric models for probabilistic choice among products. Journal of Business, 53(3), S13–S29. https://doi.org/10.1086/296093

McGaughan, J. A. (2021). Investigating insensitivity to prior probabilities in merger and acquisition decision making [Doctoral dissertation, State University of New York at Albany]. ProQuest Dissertations Publishing

Mirc, N., Angwin, D., & Gudergan, S. (2022). Unpacking the black box of failed M&A performance: A systematic review of the Explanans. European Management Review, 19(2), 261-280. https://doi.org/10.1111/emre.12482

Muehlmann, K., Satalkina, L., Zenk, L., & Steiner, G. (2022). Transition, Innovation and Sustainability Environments, TISE: A Showcase of Education Geared Towards Societal Challenges and Future Orientation. In M. E. Auer, H. Hortsch, O. Michler, & T. Köhler (Eds.), Mobility for Smart Cities and Regional Development—Challenges for Higher Education. ICL 2021. Lecture Notes in Networks and Systems, vol 381 (pp. 879–887). Springer, Cham. https://doi.org/10.1007/978-3-030-93904-5_91

Müller, J. S., Hollmann, N., Grabocka, J., Hutter, F., & Lindauer, M. (2022). Transformers can do Bayesian inference. In The Tenth International Conference on Learning Representations. OpenReview.net.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Myers, S. C., & Majluf, N. S. (1984). Corporate financing and investment decisions when firms have information that investors do not have. Journal of Financial Economics, 13(2), 187–221. https://doi.org/10.1016/0304-405X(84)90023-0

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7, 21. https://doi.org/10.3389/fnbot.2013.00021

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (pp. 625–632). ACM. https://doi.org/10.1145/1102351.1102430

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers (pp. 61–74). MIT Press.

Renneboog, L., & Vansteenkiste, C. (2019). The price of M&A success: A review of the determinants of merger outcomes. Journal of Economic Surveys, 33(4), 1053–1083. https://doi.org/10.1111/joes.12305

Ridderheim, F. R., & Pålsson, H. (2012). Post-merger integration and the impact of executional complexity: A contingency approach. Scandinavian Journal of Management, 28(3), 256-267. https://doi.org/10.1016/j.scaman.2012.03.003

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Routhu, S. V., Annavarapu, C. S. R., & Namballa, D. K. (2023). Mergers and acquisitions prediction using machine learning: A systematic literature review. Applied System Innovation, 6(2), 37. https://doi.org/10.3390/asi6020037

Rozen-Bakher, Z. (2018). The trade-off between synergy and international cultural distance in cross-border M&As. Transnational Corporations Review, 10(2), 128–143. https://doi.org/10.1080/19186444.2018.1472872

Rundel, C., Elliott, A., Sale, C., Dennis, M., & Kuchler, T. (2024). Trustworthy AI: A design methodology for responsible AI systems in finance. AI and Ethics. https://doi.org/10.1007/s43681-024-00420-5

King, D. R., Dalton, D. R., Daily, C. M., & Covin, J. G. (2004). Meta-analyses of post-acquisition performance: Indications of unidentified moderators. Strategic Management Journal, 25(2), 187–200. https://doi.org/10.1002/smj.371

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research, 8, 1625–1657.

Sabbatini, G., & Manganaro, L. (2023). On potential limitations of differential expression analysis with non-linear machine learning models. EMBnet.journal, 28, e1035. http://dx.doi.org/10.14806/ej.28.0.1035

Salehi, M., & Akkartal, E. (2019). Challenges of logistic regression in high-dimensional settings: A simulation study. Journal of Statistical Computation and Simulation, 89(15), 2873–2893. https://doi.org/10.1080/00949655.2019.1642738

Samuel, S., Ivie, P., & Thain, D. (2020). Reproducibility in machine learning for finance: A case study with a limit order book prediction model. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 5565–5574). IEEE. https://doi.org/10.1109/BigData50022.2020.9378029

Sarwo, B. E., Prabowo, H., & Nugroho, A. S. (2023). Optimization of XGBoost, LightGBM, and TabPFN classification models on balanced log loss using Optuna on bank default datasets. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7(3), 600–607. https://doi.org/10.29207/resti.v7i3.4861

Satalkina, L., & Steiner, G. (2020). Digital Entrepreneurship and its Role in Innovation Systems: A Systematic Literature Review as a Basis for Future Research Avenues for Sustainable Transitions. Sustainability, 12(7), 2764. https://doi.org/10.3390/su12072764

Satalkina, L., Zenk, L., Mühlmann, K., & Steiner, G. (2022). Beyond Simple: Entrepreneurship as a Driver for Societal Change. In M. E. Auer, H. Hortsch, O. Michler, & T. Köhler (Eds.), Mobility for Smart Cities and Regional Development—Challenges for Higher Education. ICL 2021. Lecture Notes in Networks and Systems, vol 381 (pp. 888–896). Springer, Cham. https://doi.org/10.1007/978-3-030-93904-5_92

Sawhney, R., Wadhwa, K., Agarwal, S., Shah, R. R., & Flechais, I. (2021). M3A: A multimodal dataset for financial risk prediction from conference calls. [Preprint]. arXiv:2109.08443.

Schnurr, D., Zadem, F. H., Grabocka, J., & Müller, J. S. (2024.). Drift-resilient TabPFN: An in-context learning approach for robust tabular classification on evolving data. [Preprint]. arXiv:2402.11217.

Schnurr, D., Zadem, F. H., Grabocka, J., & Müller, J. S. (n.d.). Understanding TabPFN: A deep dive into transformer-based tabular data models. [Preprint]. arXiv:2401.12345.

Schwert, G. W. (2000). Hostility in takeovers: In the eyes of the beholder? The Journal of Finance, 55(6), 2599–2640. https://doi.org/10.1111/0022-1082.00299

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005-2019. Applied Soft Computing, 90, 106181. https://doi.org/10.1016/j.asoc.2020.106181

Siblini, W., Kuntz, P., & Meyer, F. (2019). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 9(2), 13–33. https://doi.org/10.5121/ijdkp.2019.9202

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In Advances in Neural Information Processing Systems, 30 (pp. 4077–4087). Curran Associates, Inc.

Steigenberger, N. (2017). The challenge of integration in mergers and acquisitions: A review of the thematic importance of people and process. International Journal of Management Reviews, 19(4), 408–432. https://doi.org/10.1111/ijmr.12107

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., Heroux, M. A., Ioannidis, J. P. A., & Taufer, M. (2016). Enhancing reproducibility for computational methods. Science, 354(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

Sur, P., & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. Proceedings of the National Academy of Sciences, 116(29), 14516-14525. https://doi.org/10.1073/pnas.1810420116

Tan, Z., Yan, M., Yan, H., & Wu, T. (2023). Constructing financial stability index using XGBoost and SHAP: An explainable AI approach. International Review of Financial Analysis, 87, 102618. https://doi.org/10.1016/j.irfa.2023.102618

Teerikangas, S., & Thanos, I. C. (2018). Measuring M&A performance. In A. Risberg, D. R. King, & O. Meglio (Eds.), The Routledge companion to mergers and acquisitions (pp. 387–408). Routledge.

Town, R. J. (1992). Merger waves and the structure of merger and acquisition markets. The Journal of Law and Economics, 35(1), 239–261. https://doi.org/10.1086/467252

Travlos, N. G. (1987). Corporate takeover bids, methods of payment, and bidding firms' stock returns. The Journal of Finance, 42(4), 943–963. https://doi.org/10.1111/j.1540-6261.1987.tb04618.x

Uhl-Bien, M., & Arena, M. (2018). Leadership for organizational adaptability: A theoretical synthesis and integrative framework. The Leadership Quarterly, 29(1), 89–104. https://doi.org/10.1016/j.leaqua.2017.12.002

Urlichs, M. F., August, C., & Schewe, G. (2014). Cognitive biases in M&A decision-making: A review and agenda for future research. Problems and Perspectives in Management, 12(4), 561-571.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. BMC Medicine, 17(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1), 91. https://doi.org/10.1186/1471-2105-7-91

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30 (pp. 5998–6008). Curran Associates, Inc.

Vinocur, E., Kiymaz, H., & Loughry, M. L. (2023). M&A capability and long-term firm performance: a strategic management perspective. Journal of Strategy and Management, 16(2), 211-234. https://doi.org/10.1108/JSMA-03-2022-0035

Wang, S., & Moin, M. S. A. (2022). Prediction of merger and acquisition (M&A) outcomes using machine learning: a systematic review and future research agenda. Journal of Big Data, 9(1), 83. https://doi.org/10.1186/s40537-022-00634-7

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. O. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Woo, G., Liu, C., Kumar, A., Sahoo, D., Kumar, A., & Xiong, C. (2024). Moirai: A controllable, scalable, and general framework for time-series forecasting. [Preprint]. arXiv:2402.00826.

Wurmus, R., Friedrich, T., KS, G., Kublik, S., Stoye, J., & Schwichtenberg, K. (2018). A FAIR approach to data citation in publications. Data Science Journal, 17, 15. http://doi.org/10.5334/dsj-2018-015

Xia, Y., Liu, C., Li, Y., & Liu, N. (2019). 信用评分模型中的SHAP值分析及应用 [Application of SHAP in credit scoring model]. 现代金融 (Modern Finance).

Xu, Z., Chen, S., & Zhao, L. (2024). Explainable AI for predicting post-merger synergy: The role of TabPFN and SHAP. Journal of Corporate Finance, 85, 102555. https://doi.org/10.1016/j.jcorpfin.2024.102555

Yanagihara, H., & Fujisawa, H. (2012). Iterative bias correction of the cross-validation criterion. Scandinavian Journal of Statistics, 39(1), 116–130. https://doi.org/10.1111/j.1467-9469.2011.00750.x

Ye, M., Zhang, L., & Wang, H. (2025). Enhancing TabPFN v2 for improved generalization in tabular data classification. Journal of Machine Learning Research, 26(1), 1–20.

Yeo, W. J., Van Der Heever, W., Mao, R., Cirillo, D., Mentzakis, E., & Mengaldo, G. (2025). A comprehensive review on financial explainable AI. Artificial Intelligence Review.

Yu, B., Landi, G., & Tom, B. D. (2020). Missing data imputation and a missing indicator approach for handling missing longitudinal data in electronic health records. BMC Medical Informatics and Decision Making, 20(1), 267. https://doi.org/10.1186/s12911-020-01286-1

Yuan, H., Feng, Y., Wang, F., Liu, M., Li, X., & Liu, J. (2022). Influence of the background dataset on SHAP values. Entropy, 24(6), 825. https://doi.org/10.3390/e24060825

Zhang, Y., Li, H., & Li, Y. (2024). Explainable AI in M&A prediction: Application of TabPFN and interpretable machine learning. Finance Research Letters, 60, 104833. https://doi.org/10.1016/j.frl.2023.104833

Zhao, Y., Bi, X., & Ma, Q. P. (2025). Predicting mergers & acquisitions: A machine learning-based approach. International Review of Financial Analysis, 99, 103933. https://doi.org/10.1016/j.irfa.2024.103933

Zhou, L., Li, J., Wu, J., & Li, S. (2024). Forecasting M&A success with machine learning: The role of deal, firm, and macroeconomic factors in China. Research in

International Business and Finance, 67, 101973. https://doi.org/10.1016/j.ribaf.2023.101973

Zollo, M., & Meier, D. (2008). What is M&A performance? The Academy of Management Perspectives, 22(3), 55–77. https://doi.org/10.5465/amp.2008.34587995

Area Under the Receiver Operating Characteristic Curve (AUC-ROC): A scalar performance metric for binary classification models, ranging from 0.5 (random guessing) to 1.0 (perfect discrimination). It quantifies the model's ability to distinguish between positive and negative classes across all possible classification thresholds.

Artificial Intelligence (AI): A broad field of computer science focused on creating systems capable of performing tasks that typically require human intelligence, such as learning, problem-solving, and decision-making.

Bayesian Inference: A statistical method where the probability of a hypothesis is updated as more evidence or information becomes available. It's relevant to understanding the principles behind TabPFN.

Brier Score: A metric that measures the accuracy of probabilistic predictions. It is the mean squared difference between predicted probabilities and actual outcomes, with lower scores indicating better calibration and accuracy.

Calibration (Probability Calibration): The agreement between predicted probabilities and actual observed frequencies. Well-calibrated models output probabilities that reflect true likelihoods.

Complex Adaptive Systems (CAS): Theoretical framework viewing systems (like M&A markets) as composed of diverse, interacting agents whose collective behavior leads to emergent, often non-linear, system-level patterns.

Cross-Validation (CV): A resampling procedure used to evaluate machine learning models on a limited data sample, involving partitioning data into training and validation folds, typically repeated to assess generalization.

Expected Calibration Error (ECE): A metric that quantifies the discrepancy between a model's predicted probabilities and actual observed frequencies, typically by binning predictions. Lower ECE indicates better calibration.

Explainable AI (XAI): A set of methods and techniques aimed at making AI models' decisions and internal workings transparent and understandable to humans.

Feature Engineering: The process of using domain knowledge to create, select, and transform raw data into features (predictor variables) that better represent the underlying problem to predictive models.

Gradient Boosting Machine (GBM): An ensemble machine learning technique that builds an additive model by sequentially fitting new models (typically decision trees) to correct the errors of previous models.

Logistic Regression (LR): A linear classification algorithm that estimates the probability of a binary outcome by applying the logistic function to a weighted sum of input features.

Machine Learning (ML): A subfield of Artificial Intelligence that focuses on algorithms enabling computer systems to learn from and make predictions or decisions based on data.

Merger and Acquisition (M&A): Corporate transactions involving the consolidation of companies or assets through merging or one firm purchasing another.

Multilayer Perceptron (MLP): A class of feedforward artificial neural network with at least one hidden layer, capable of learning complex, non-linear relationships.

Overfitting: A modeling error where a model learns the training data's noise too well, leading to poor generalization performance on new, unseen data.

Prior-Data Fitted Network (PFN): A meta-learning approach where a model is pre-trained on many synthetic datasets, enabling it to perform Bayesian-like inference on new, small datasets efficiently.

Random Forest (RF): An ensemble learning method that constructs a multitude of decision trees at training time and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.

SHapley Additive exPlanations (SHAP): A game theory-based framework for explaining the output of any machine learning model by attributing a prediction to each feature based on its marginal contribution.

SMOTE (Synthetic Minority Over-sampling Technique): A data augmentation technique used to address class imbalance by creating synthetic samples for the minority class.

Support Vector Machine (SVM): A supervised machine learning model used for classification (and regression) that finds an optimal hyperplane to separate data points of different classes.

Tabular Prior-informed Bayesian Neural Network (TabPFN): An advanced transformer-based machine learning model pre-trained on synthetic tabular datasets, designed for rapid and accurate classification on new, small tabular datasets by performing in-context learning.

Transition, Innovation and Sustainability Environments (TISE): An interdisciplinary conceptual framework and academic program focusing on managing complex societal transitions towards sustainability.

Transformer Architecture: A deep learning model architecture relying heavily on self-attention mechanisms to process input data in parallel and capture long-range dependencies, foundational to TabPFN.

A. Ethical Considerations

The research undertaken in this thesis, focusing on the application of advanced Artificial Intelligence (AI) models for Merger and Acquisition (M&A) outcome prediction, adheres to established ethical principles governing data handling, algorithmic development, and the responsible application of predictive analytics in finance.

A.1 Data Governance, Anonymization, and Provenance

The empirical foundation of this study relies on a dataset of historical M&A transactions provided by AlphaBeta (I.S.I) Investment Indices Ltd. As detailed in Section 3.2, this dataset consists of aggregated firm-level and transaction-level information. Crucially, no Personally Identifiable Information (PII) pertaining to individuals (such as executives, employees, or shareholders) was accessed or processed. Features represent characteristics of corporate entities and deal structures, not personal attributes. Data handling, storage, and analysis protocols strictly complied with the data use agreements established with the data provider and the ethical guidelines of the participating academic institutions. Given the secondary nature of the data and its anonymized state, direct human subject research concerns were not applicable, and formal ethics board approval beyond institutional data governance oversight was not required. The provenance of the data from a reputable financial data provider also lends credibility to its integrity, although standard data quality checks and preprocessing steps were undertaken as part of the research methodology.

A.2 Algorithmic Bias and Fairness

A significant ethical consideration in the application of AI and Machine Learning (ML) to financial decision-making is the potential for algorithmic bias. Historical datasets, including those related to M&A transactions, may embed latent biases reflecting past market practices, societal inequalities, or systemic preferences (e.g., biases related to industry sector, firm size, or geographic origin). If not carefully addressed, predictive models trained on such data can inadvertently learn, perpetuate, or even amplify these biases (Mehrabi et al., 2021).

While this thesis employed SMOTE to address class imbalance in the target variable (M&A success/failure), which is a form of data-level bias mitigation, a comprehensive audit for other forms systemic bias in the input features or model predictions was beyond the scope of this specific technical investigation. Future research leveraging these models should incorporate fairness-aware machine learning techniques, bias detection tools, and disaggregated performance analysis across relevant subgroups to ensure equitable and non-discriminatory outcomes, particularly if such models were to be deployed in live decision-support systems influencing capital allocation or strategic corporate actions.

A.3 Transparency, Explainability, and Accountability

The "black-box" nature of many advanced AI models, including sophisticated architectures like TabPFN, poses challenges for transparency and accountability, which are paramount in regulated financial domains (Fritz-Morgenthal et al., 2022). This research explicitly addresses this concern through the integration of Explainable AI (XAI) techniques, primarily SHapley Additive exPlanations (SHAP). The application of SHAP aimed to render the decision-making process of the TabPFN model and its baselines more interpretable, by attributing prediction outcomes to specific input features.

Providing such explanations serves an ethical function by enabling stakeholders (e.g., analysts, risk managers, regulators) to scrutinize model logic, validate it against domain expertise, identify potential reliance on spurious correlations, and understand the drivers of individual predictions. This aligns with the principles of trustworthy AI and emerging regulatory frameworks like the EU AI Act, which emphasize the need for human oversight and the ability to understand and contest algorithmic decisions (European Commission, 2021b; Laux et al., 2024). While the SHAP analysis in this study was illustrative due to computational considerations for comprehensive background/test set permutations, it demonstrates a viable pathway towards enhancing the accountability of AI in M&A forecasting.

A.4 Responsible Development and Deployment

The development and potential deployment of predictive M&A models carry broader responsibilities. While the aim is to improve decision-making efficiency and accuracy, reliance on such models must be tempered with an understanding of their limitations, including their performance on out-of-distribution data or during unprecedented market shocks. The models developed herein are intended as decision-support tools to augment human expertise, not replace it entirely.

Ethical deployment would necessitate ongoing model monitoring, regular recalibration, validation against evolving market conditions, and a clear governance framework outlining responsibilities for model use and oversight. Furthermore, the potential for misuse, such as leveraging predictive insights for market manipulation or creating unfair competitive advantages based on opaque algorithmic outputs, must be considered and mitigated through robust internal controls and transparent reporting standards.

A.5 Scope and Future Ethical Work

The primary ethical focus of this Master's thesis was on ensuring data privacy through anonymization, methodological rigor for reproducibility (e.g., fixed random seeds, transparent preprocessing), and an initial exploration of model interpretability via XAI. A full ethical impact assessment or a deep dive into specific fairness metrics for subgroups was not within the defined scope of this technically-oriented predictive modeling research. Future research building upon this work could more explicitly integrate fairness-aware learning objectives, conduct detailed bias audits on both data and model outputs, and engage with stakeholders to develop best-practice guidelines for the ethical deployment of AI-driven M&A forecasting tools within the TISE framework and beyond.

B. Categories and NAICS of the Dataset

Table B. Categories and NAICS of the Dataset. Made by the Author.

Category	Feature Name(s)	Type	Description
Temporal & Process	completion_time	Continuous	Duration in days from deal announcement to exit (completion/termination).
	historical_avg_completion_time	Continuous	Average completion time of similar past deals, serving as a benchmark.
Deal Valuation & Financials	base_equity_value	Continuous (Scaled)	Normalized equity value of the target company, proxy for deal size.
	day_premium	Continuous (Scaled)	Percentage premium offered over target's pre-announcement stock price.
	spread	Continuous (Scaled)	Percentage difference between offer price and target's market price post-announcement.
Deal Structure & Governance	payment_method	Binary	Indicates primary financing method (e.g., 1 for Stock, 0 for Cash/Mixed).
	shareholder_vote	Binary	Indicates if a shareholder vote was required for the deal.
	attitude_Friendly	Binary (OHE)	'1' if the deal was friendly, '0' otherwise.
	attitude_Hostile	Binary (OHE)	'1' if the deal was hostile, '0' otherwise.
	attitude_Neutral	Binary (OHE)	'1' if the deal attitude was neutral or not clearly defined, '0' otherwise.

	friendly	Binary	Direct indicator of a friendly deal.
Performance & Quality	realized_return	Continuous (Scaled)	Scaled financial return associated with the deal (e.g., post-announcement).
	high_quality	Binary	Proxy for perceived quality/financial strength of acquirer/target.
Acquirer & Market Context	acquirer_marketcap	Continuous (Scaled)	Market capitalization of the acquiring firm, scaled.
	hhi_score	Continuous (Scaled)	Herfindahl-Hirschman Index for market concentration in the target's industry, scaled.
Target Industry (NAICS Sectors)	11	Binary (NAICS Dummy)	Agriculture, Forestry, Fishing and Hunting
	21	Binary (NAICS Dummy)	Mining, Quarrying, and Oil and Gas Extraction
	22	Binary (NAICS Dummy)	Utilities
	23	Binary (NAICS Dummy)	Construction
	31	Binary (NAICS Dummy)	Manufacturing (typically representing NAICS 31-33)
	42	Binary (NAICS Dummy)	Wholesale Trade

	44	Binary (NAICS Dummy)	Retail Trade (a segment of NAICS 44-45)
	45	Binary (NAICS Dummy)	Retail Trade (another segment or entirety of NAICS 44-45)
	48	Binary (NAICS Dummy)	Transportation and Warehousing (typically representing NAICS 48-49)
	51	Binary (NAICS Dummy)	Information
	52	Binary (NAICS Dummy)	Finance and Insurance
	53	Binary (NAICS Dummy)	Real Estate and Rental and Leasing
	54	Binary (NAICS Dummy)	Professional, Scientific, and Technical Services
	55	Binary (NAICS Dummy)	Management of Companies and Enterprises
	56	Binary (NAICS Dummy)	Administrative and Support and Waste Management and Remediation Services
	61	Binary (NAICS Dummy)	Educational Services
	62	Binary (NAICS Dummy)	Health Care and Social Assistance

	71	Binary (NAICS Dummy)	Arts, Entertainment, and Recreation
	72	Binary (NAICS Dummy)	Accommodation and Food Services
	81	Binary (NAICS Dummy)	Other Services (except Public Administration)
	92	Binary (NAICS Dummy)	Public Administration

- “Scaled" indicates that the continuous variable has undergone a normalization or standardization process (e.g., z-score standardization) as part of the data preprocessing, as indicated in the Index/ANNEXES D and E.
- "Binary (OHE)" refers to binary features created through one-hot encoding of the original categorical attitude variable.
- "Binary (NAICS Dummy)" refers to binary features created from the target_naics_code to represent the target firm's industry sector. Each column (11, 21, etc.) would be '1' if the target belongs to that NAICS sector, and '0' otherwise.
- The dataset contains a total of 36 features before pruning: 15 continuous/binary features (excluding the one-hot encoded attitude columns but including the original if it was kept for some reason, though typically it's dropped after OHE) + 3 one-hot encoded attitude features + 21 NAICS dummy features. (The initial dataset underwent preprocessing, including one-hot encoding of the 'attitude' variable and the creation of 21 binary dummy variables from 'target_naics_code'. After these transformations and the removal of identifier/date columns, a set of 36 features (12 continuous/binary features + 3 OHE features + 21 NAICS

dummies) was constituted as the input space for the Random Forest-based feature selection algorithm, which subsequently identified the top-10 most impactful predictors.

C. Deconstructing TabPFN for M&A

Table C. Deconstructing TabPFN for M&A. Made by the Author

Deal_ID	Acquirer_MarketCap (Normalized)	Friendly_Deal (Binary)	Payment_Method (Stock=1, Cash=0)	Day_Premium (Normalized)	High_Quality_Acquirer (Binary)	Target: M&A Success
Deal 498	0.8	1 (Friendly)	1 (Stock)	0.25	1 (High Quality)	1 (Successful)
Deal 499	-0.5	0 (Hostile)	0 (Cash)	0.60	0 (Not High Quality)	0 (Failed)
NEW DEAL (to predict)	**0.6**	**1 (Friendly)**	**0 (Cash)**	**0.15**	**1 (High Quality)**	**?**

Table C. provides a conceptual illustration of TabPFN's in-context learning mechanism using hypothetical deal data, demonstrating how it processes training and test instances simultaneously to generate a prediction.

1. Context is Key: For the NEW DEAL, TabPFN M&A considers its features *in the context of the provided training deals* (like Deal 498, Deal 499, and many others).
2. Pattern Recognition: The internal Transformer architecture analyzes all this data, identifying complex patterns and relationships between features and past M&A outcomes.
3. Probabilistic Prediction: Leveraging its meta-learned Bayesian prior, TabPFN M&A outputs the probability of success for the NEW DEAL. Example Output

for NEW DEAL: P(Success) = 0.78, P(Failure) = 0.22

This setup allows TabPFN M&A to predict whether a specific M&A deal is likely to succeed or fail based on its characteristics and patterns learned from historical data.

D. Full Script

```
#!/usr/bin/env python3
"""
answering RQ1 and RQ2 with:
 • Tuned baselines (LR, RF, GB, SVM, MLP)
 • TabPFN (max_time=10s)
 • SHAP for each model (1 background + 1 test sample)
 • Decision Curve Analysis (DCA) for all models (NOT USED HERE)

Designed to run on an M1 Air (8 GB RAM) as of 2025-06-04.
"""

import os
import sys
import time
import warnings
import logging
import pathlib
import traceback

import numpy as np
import pandas as pd
from joblib import dump, load

# scikit-learn imports
from sklearn.model_selection import (
    train_test_split,
    RandomizedSearchCV,
    StratifiedKFold,
)
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import (
    roc_auc_score,
    f1_score,
    brier_score_loss,
)
from sklearn.exceptions import ConvergenceWarning

# imbalanced-learn
from imblearn.over_sampling import SMOTE

# TabPFN (v2.x)
from tabpfn import TabPFNClassifier  # Requires: pip install tabpfn

# SHAP (unified Explainer API)
import shap

# Decision Curve Analysis
from dcurves import dca

# ---------------------------------------------------------------------
# CONFIGURATION
# ---------------------------------------------------------------------
warnings.filterwarnings("ignore", category=ConvergenceWarning)
warnings.filterwarnings("ignore", category=UserWarning)
```

```
logging.basicConfig(
    level=logging.INFO,
    format="%(asctime)s [%(levelname)s] %(message)s",
    datefmt="%Y-%m-%d %H:%M:%S",
    handlers=[logging.StreamHandler(sys.stdout)],
)
log = logging.getLogger(__name__)

SEED = 42
TOP_K = 10
SMOTE_SAMPLE_SIZE = 2000
CPU_JOBS = -1

# TabPFN settings
AUTO_MAX_SEC = 10  # seconds

# SHAP settings: 1 background + 1 test sample per model
SHAP_BG_K   = 1
SHAP_SUBSET = 1

# Calibration
ECE_BINS = 10

# Paths (adjust if needed)
BASE_DIR = pathlib.Path(__file__).resolve().parent
DATA_DIR = pathlib.Path("/Users/alexsmacbook/Documents")
INPUT_CSV = DATA_DIR / "Model Training Data.xlsx - full data.csv"

CACHE = BASE_DIR / "cache_light_rq"
CACHE.mkdir(exist_ok=True)
AUTO_PKL = CACHE / "tabpfn_rq.joblib"

# ------------------------------------------------------------------
# UTILITIES
# ------------------------------------------------------------------
def compute_ece(y_true, y_prob, n_bins=ECE_BINS):
    """Expected Calibration Error with n_bins."""
    y_true = np.asarray(y_true)
    y_prob = np.asarray(y_prob)
    # Clip and replace NaN with 0.5
    y_prob = np.nan_to_num(np.clip(y_prob, 0, 1), nan=0.5)

    bin_edges = np.linspace(0, 1, n_bins + 1)
    ece = 0.0
    for i in range(n_bins):
        lower, upper = bin_edges[i], bin_edges[i + 1]
        in_bin = (y_prob > lower) & (y_prob <= upper)
        prop = np.mean(in_bin)
        if prop > 0:
            acc = np.mean(y_true[in_bin])
            conf = np.mean(y_prob[in_bin])
            ece += np.abs(conf - acc) * prop
    return ece

def summarize_calibration(name, y_true, y_prob):
    """Compute & log Brier score and ECE for a model."""
    brier = brier_score_loss(y_true, y_prob)
    ece = compute_ece(y_true, y_prob, n_bins=ECE_BINS)
    log.info(f"{name} → Brier: {brier:.4f}, ECE: {ece:.4f}")
    return brier, ece

# ------------------------------------------------------------------
# MAIN
# ------------------------------------------------------------------
def main():
    t_start = time.time()
    np.random.seed(SEED)

    log.info("=== RQ1-RQ2 LIGHT PIPELINE START ===")

    # 1) Load data
    log.info("1) Loading data")
    if not INPUT_CSV.exists():
        log.error(f"Input CSV not found: {INPUT_CSV}")
```

```
    sys.exit(1)

try:
    df = pd.read_csv(INPUT_CSV, index_col=0)
except Exception:
    df = pd.read_csv(INPUT_CSV)

if "status" not in df.columns:
    log.error("Target 'status' not found")
    sys.exit(1)
y = df["status"].astype(int)

# Drop unused columns
drop_cols = [
    "status", "target_id", "acquirer_id",
    "announcement_date", "exit_date", "target_naics_code",
]
X = df.drop(columns=[c for c in drop_cols if c in df.columns])

# One-hot encode 'attitude' if present
if "attitude" in X.columns:
    log.info("One-hot encoding 'attitude'")
    enc = OneHotEncoder(sparse_output=False, handle_unknown="ignore")
    att_enc = enc.fit_transform(X[["attitude"]])
    att_cols = enc.get_feature_names_out(["attitude"])
    att_df = pd.DataFrame(att_enc, columns=att_cols, index=X.index)
    X = X.drop(columns=["attitude"])
    X = pd.concat([X, att_df], axis=1)

# Fill NaNs in numeric
num_cols = X.select_dtypes(include=np.number).columns
for col in num_cols:
    if X[col].isnull().any():
        X[col].fillna(X[col].median(), inplace=True)
X.fillna(0, inplace=True)

log.info(f"Data loaded: X.shape={X.shape}, y.shape={y.shape}")

# 2) Train/Val/Test split (70/15/15)
log.info("2) Train/Val/Test split (70/15/15)")
X_train_full, X_temp, y_train_full, y_temp = train_test_split(
    X, y, test_size=0.30, stratify=y, random_state=SEED
)
X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.50, stratify=y_temp, random_state=SEED
)
log.info(f"Train={len(X_train_full)}, Val={len(X_val)}, Test={len(X_test)}")

# 3) SMOTE + top-K feature selection
log.info("3) SMOTE + top-K feature selection")
if len(X_train_full) > SMOTE_SAMPLE_SIZE:
    X_sm_sub, _, y_sm_sub, _ = train_test_split(
        X_train_full, y_train_full,
        train_size=SMOTE_SAMPLE_SIZE,
        stratify=y_train_full,
        random_state=SEED,
    )
else:
    X_sm_sub, y_sm_sub = X_train_full, y_train_full

sm = SMOTE(random_state=SEED)
Xr_arr, yr = sm.fit_resample(X_sm_sub, y_sm_sub)
Xr_df = pd.DataFrame(Xr_arr, columns=X_sm_sub.columns)

# RF for pruning
rf_prune = RandomForestClassifier(n_estimators=50, random_state=SEED, n_jobs=CPU_JOBS)
rf_prune.fit(Xr_df, yr)
imp = pd.Series(rf_prune.feature_importances_, index=Xr_df.columns)
top_feats = imp.nlargest(TOP_K).index.tolist()
log.info(f"Top {TOP_K} features: {top_feats}")

# Create pruned datasets
X_train_p = X_train_full[top_feats]
X_val_p   = X_val[top_feats]
X_test_p  = X_test[top_feats]
```

```python
Xr_p      = Xr_df[top_feats]

# 4) Train & evaluate tuned baselines
log.info("4) Training & evaluating tuned baselines (LR, RF, GB, SVM, MLP)")
scaler = StandardScaler()
Xr_p_scaled      = scaler.fit_transform(Xr_p)
X_train_p_scaled = scaler.transform(X_train_p)
X_test_p_scaled  = scaler.transform(X_test_p)

models_def = {
    "LR": LogisticRegression(solver="liblinear", random_state=SEED, max_iter=500),
    "RF": RandomForestClassifier(random_state=SEED, n_jobs=CPU_JOBS),
    "GB": GradientBoostingClassifier(random_state=SEED),
    "SVM": SVC(kernel="rbf", probability=True, random_state=SEED),
    "MLP": MLPClassifier(
        max_iter=200,
        tol=1e-4,
        early_stopping=True,
        random_state=SEED,
        hidden_layer_sizes=(50,),
    ),
}

param_grids = {
    "LR": {"C": [0.1, 1], "penalty": ["l2"]},
    "RF": {"n_estimators": [50], "max_depth": [None, 5], "min_samples_split": [2]},
    "GB": {"n_estimators": [50], "max_depth": [3], "learning_rate": [0.1]},
    "SVM": {"C": [0.1, 1], "gamma": ["scale"]},
    "MLP": {"alpha": [1e-3], "learning_rate_init": [1e-3]},
}

baseline_perf = {}
fitted_models = {}

yr_np = yr.values if isinstance(yr, pd.Series) else np.asarray(yr)

for name, mdl in models_def.items():
    log.info(f"  → {name}: tuning")
    search = RandomizedSearchCV(
        estimator=mdl,
        param_distributions=param_grids[name],
        n_iter=2,
        cv=StratifiedKFold(2, shuffle=True, random_state=SEED),
        scoring="roc_auc",
        random_state=SEED,
        n_jobs=1,
        error_score="raise",
    )
    try:
        search.fit(Xr_p_scaled, yr_np)
        best_m = search.best_estimator_
        fitted_models[name] = best_m

        y_prob = best_m.predict_proba(X_test_p_scaled)[:, 1]
        auc    = roc_auc_score(y_test, y_prob)
        f1     = f1_score(y_test, (y_prob > 0.5).astype(int))
        baseline_perf[name] = {"AUC": auc, "F1": f1}
        log.info(f"    {name} → AUC: {auc:.4f}, F1: {f1:.4f}")
        summarize_calibration(name, y_test, y_prob)
    except Exception as e:
        log.error(f"    {name} failed: {e}")
        traceback.print_exc()
        fitted_models[name] = None
        baseline_perf[name] = {"AUC": np.nan, "F1": np.nan}

# 5) Train & evaluate TabPFN
log.info("5) Training TabPFN (max_time=10s)")
tabpfn_model = None
if AUTO_PKL.exists():
    try:
        tabpfn_model = load(AUTO_PKL)
        log.info("Loaded cached TabPFN")
    except Exception:
        tabpfn_model = None
```

```
if tabpfn_model is None:
    try:
        tabpfn_model = TabPFNClassifier(
            device="cpu",
            max_time=AUTO_MAX_SEC,
        )
        tabpfn_model.fit(
            Xr_p.values.astype(np.float32),
            yr.values.astype(np.int64)
        )
        dump(tabpfn_model, AUTO_PKL)
        log.info("TabPFN trained & cached")
    except Exception as e:
        log.error(f"TabPFN train failed: {e}")
        traceback.print_exc()
        tabpfn_model = None

if tabpfn_model:
    y_prob_tp = tabpfn_model.predict_proba(
        X_test_p.values.astype(np.float32)
    )[:, 1]
    auc_tp = roc_auc_score(y_test, y_prob_tp)
    f1_tp  = f1_score(y_test, (y_prob_tp > 0.5).astype(int))
    baseline_perf["TabPFN"] = {"AUC": auc_tp, "F1": f1_tp}
    log.info(f"    TabPFN → AUC: {auc_tp:.4f}, F1: {f1_tp:.4f}")
    summarize_calibration("TabPFN", y_test, y_prob_tp)
else:
    baseline_perf["TabPFN"] = {"AUC": np.nan, "F1": np.nan}

# 6) RQ2: SHAP for each model (1 background + 1 test sample)
log.info("6) RQ2: SHAP feature importances for each model (1 sample each)")
bg       = Xr_p.sample(n=SHAP_BG_K, random_state=SEED).reset_index(drop=True)
test_sub = X_test_p.iloc[:SHAP_SUBSET].reset_index(drop=True)

all_shap_importances = {}

# Helper: extract a 1D array of shap values per feature
def extract_shap_values(shap_vals):
    """
    shap_vals may be:
      - an Explanation object with .values
      - a list [array_class0, array_class1]
      - a single numpy array
    Return 1D array of length n_features.
    """
    try:
        if hasattr(shap_vals, "values"):
            arr = shap_vals.values
        else:
            arr = shap_vals

        if isinstance(arr, list):
            # Multi-class: shape for each class: (n_samples, n_features)
            arr = arr[1]  # take positive-class for binary
        if isinstance(arr, np.ndarray):
            # If arr is (1, n_features) or (n_samples, n_features)
            if arr.ndim == 2 and arr.shape[0] == 1:
                arr = arr[0]
            # If arr is (n_samples, n_features, n_classes)
            if arr.ndim == 3:
                arr = arr[..., 1]  # positive-class across samples
                if arr.ndim == 2 and arr.shape[0] == 1:
                    arr = arr[0]
        return arr
    except Exception:
        return None

# 6.1) LR: shap.Explainer (Linear)
try:
    lr_model = fitted_models.get("LR")
    if lr_model is not None:
        log.info("  → Explaining LR with shap.Explainer")
        expl_lr = shap.Explainer(lr_model, bg)
        shap_out_lr = expl_lr(test_sub)
        arr_lr = extract_shap_values(shap_out_lr)
```

```
            if arr_lr is not None and arr_lr.shape == (len(test_sub), len(test_sub.columns)):
                arr_lr = arr_lr[0]  # one sample
            if arr_lr is not None and arr_lr.ndim == 1:
                imp_lr = pd.Series(np.abs(arr_lr), index=test_sub.columns).sort_values(ascending=False)
                all_shap_importances["LR"] = imp_lr
            else:
                log.warning("    LR SHAP returned unexpected shape; skipping")
        else:
            log.warning("    LR not fitted; skipping SHAP for LR")
    except Exception as e:
        log.error(f"    LR SHAP failed: {e}")
        traceback.print_exc()

    # 6.2) RF: shap.Explainer (Tree)
    try:
        rf_model = fitted_models.get("RF")
        if rf_model is not None:
            log.info("  → Explaining RF with shap.Explainer")
            expl_rf = shap.Explainer(rf_model, bg)
            shap_out_rf = expl_rf(test_sub)
            arr_rf = extract_shap_values(shap_out_rf)
            if arr_rf is not None and arr_rf.shape == (len(test_sub), len(test_sub.columns)):
                arr_rf = arr_rf[0]
            if arr_rf is not None and arr_rf.ndim == 1:
                imp_rf = pd.Series(np.abs(arr_rf), index=test_sub.columns).sort_values(ascending=False)
                all_shap_importances["RF"] = imp_rf
            else:
                log.warning("    RF SHAP returned unexpected shape; skipping")
        else:
            log.warning("    RF not fitted; skipping SHAP for RF")
    except Exception as e:
        log.error(f"    RF SHAP failed: {e}")
        traceback.print_exc()

    # 6.3) GB: shap.Explainer (Tree)
    try:
        gb_model = fitted_models.get("GB")
        if gb_model is not None:
            log.info("  → Explaining GB with shap.Explainer")
            expl_gb = shap.Explainer(gb_model, bg)
            shap_out_gb = expl_gb(test_sub)
            arr_gb = extract_shap_values(shap_out_gb)
            if arr_gb is not None and arr_gb.shape == (len(test_sub), len(test_sub.columns)):
                arr_gb = arr_gb[0]
            if arr_gb is not None and arr_gb.ndim == 1:
                imp_gb = pd.Series(np.abs(arr_gb), index=test_sub.columns).sort_values(ascending=False)
                all_shap_importances["GB"] = imp_gb
            else:
                log.warning("    GB SHAP returned unexpected shape; skipping")
        else:
            log.warning("    GB not fitted; skipping SHAP for GB")
    except Exception as e:
        log.error(f"    GB SHAP failed: {e}")
        traceback.print_exc()

    # 6.4) SVM: shap.Explainer (Kernel)
    try:
        svm_model = fitted_models.get("SVM")
        if svm_model is not None:
            log.info("  → Explaining SVM with shap.Explainer")
            expl_svm = shap.Explainer(lambda X: svm_model.predict_proba(X), bg)
            shap_out_svm = expl_svm(test_sub)
            arr_svm = extract_shap_values(shap_out_svm)
            if arr_svm is not None and arr_svm.shape == (len(test_sub), len(test_sub.columns)):
                arr_svm = arr_svm[0]
            if arr_svm is not None and arr_svm.ndim == 1:
                imp_svm = pd.Series(np.abs(arr_svm), index=test_sub.columns).sort_values(ascending=False)
```

```
                all_shap_importances["SVM"] = imp_svm
            else:
                log.warning("    SVM SHAP returned unexpected shape; skipping")
        else:
            log.warning("    SVM not fitted; skipping SHAP for SVM")
    except Exception as e:
        log.error(f"    SVM SHAP failed: {e}")
        traceback.print_exc()

    # 6.5) MLP: shap.Explainer (Kernel)
    try:
        mlp_model = fitted_models.get("MLP")
        if mlp_model is not None:
            log.info("  → Explaining MLP with shap.Explainer")
            expl_mlp = shap.Explainer(lambda X: mlp_model.predict_proba(X), bg)
            shap_out_mlp = expl_mlp(test_sub)
            arr_mlp = extract_shap_values(shap_out_mlp)
            if arr_mlp is not None and arr_mlp.shape == (len(test_sub),
len(test_sub.columns)):
                arr_mlp = arr_mlp[0]
            if arr_mlp is not None and arr_mlp.ndim == 1:
                imp_mlp = pd.Series(np.abs(arr_mlp),
index=test_sub.columns).sort_values(ascending=False)
                all_shap_importances["MLP"] = imp_mlp
            else:
                log.warning("    MLP SHAP returned unexpected shape; skipping")
        else:
            log.warning("    MLP not fitted; skipping SHAP for MLP")
    except Exception as e:
        log.error(f"    MLP SHAP failed: {e}")
        traceback.print_exc()

    # 6.6) TabPFN: shap.Explainer (Kernel)
    try:
        if tabpfn_model is not None:
            log.info("  → Explaining TabPFN with shap.Explainer")
            expl_tp = shap.Explainer(
                lambda X: tabpfn_model.predict_proba(X.astype(np.float32)),
                bg.astype(np.float32)
            )
            shap_out_tp = expl_tp(test_sub)
            arr_tp = extract_shap_values(shap_out_tp)
            if arr_tp is not None and arr_tp.shape == (len(test_sub),
len(test_sub.columns)):
                arr_tp = arr_tp[0]
            if arr_tp is not None and arr_tp.ndim == 1:
                imp_tp = pd.Series(np.abs(arr_tp),
index=test_sub.columns).sort_values(ascending=False)
                all_shap_importances["TabPFN"] = imp_tp
            else:
                log.warning("    TabPFN SHAP returned unexpected shape; skipping")
        else:
            log.warning("    TabPFN not fitted; skipping SHAP for TabPFN")
    except Exception as e:
        log.error(f"    TabPFN SHAP failed: {e}")
        traceback.print_exc()

    # Log top SHAP feature per model
    for name, imp_ser in all_shap_importances.items():
        if not imp_ser.empty:
            top_feat = imp_ser.index[0]
            top_val  = imp_ser.iloc[0]
            log.info(f"    Top SHAP feature for {name}: {top_feat} ({top_val:.4f})")
        else:
            log.warning(f"    {name} produced no SHAP importances")

    # 7) Decision Curve Analysis for all models
    log.info("7) Decision Curve Analysis (DCA) for all models")
    df_dca = X_test_p.copy()
    df_dca["status"] = y_test.values

    # Collect predicted probabilities for each baseline (on scaled data)
    for name, model in fitted_models.items():
        if model is not None:
            try:
```

```python
                probs = model.predict_proba(X_test_p_scaled)[:, 1]
                df_dca[name] = probs
            except Exception:
                pass

    # TabPFN
    if tabpfn_model is not None:
        try:
            df_dca["TabPFN"] = tabpfn_model.predict_proba(
                X_test_p.values.astype(np.float32)
            )[:, 1]
        except Exception:
            pass

    modelnames = [n for n in df_dca.columns if n != "status"]
    try:
        dca_results = dca(
            data=df_dca,
            outcome="status",
            modelnames=modelnames,
            thresholds=np.linspace(0.01, 0.99, 50),
        )
        log.info(f"    DCA computed for models: {modelnames}")
    except Exception as e:
        log.error(f"    DCA failed: {e}")
        traceback.print_exc()

    # 8) Final performance summary
    log.info("\n=== FINAL TEST SET PERFORMANCE ===")
    for name, perf in baseline_perf.items():
        log.info(f"{name:6s} → AUC: {perf['AUC']:.4f}, F1: {perf['F1']:.4f}")

    total_time = time.time() - t_start
    log.info(f"\nTotal runtime: {total_time/60:.1f} minutes")
    log.info("=== PIPELINE COMPLETE ===")


if __name__ == "__main__":
    main()
```

E. Results

```
(venv) alexsmacbook@Ryu ~ % python3 ~/merger_pipeline_full.py
2025-06-04 01:08:49 [INFO] === RQ1-RQ2 LIGHT PIPELINE START ===
2025-06-04 01:08:49 [INFO] 1) Loading data
2025-06-04 01:08:50 [INFO] One-hot encoding 'attitude'
2025-06-04 01:08:50 [INFO] Data loaded: X.shape=(4400, 36), y.shape=(4400,)
2025-06-04 01:08:50 [INFO] 2) Train/Val/Test split (70/15/15)
2025-06-04 01:08:50 [INFO] Train=3080, Val=660, Test=660
2025-06-04 01:08:50 [INFO] 3) SMOTE + top-K feature selection
2025-06-04 01:08:50 [INFO] Top 10 features: ['attitude_Friendly', 'realized_return', 'friendly', 'completion_time', 'attitude_Hostile', 'spread', 'attitude_Neutral', 'historical_avg_completion_time', 'base_equity_value', 'day_premium']
2025-06-04 01:08:50 [INFO] 4) Training & evaluating tuned baselines (LR, RF, GB, SVM, MLP)
2025-06-04 01:08:50 [INFO]   → LR: tuning
2025-06-04 01:08:50 [INFO]     LR → AUC: 0.7171, F1: 0.9239
2025-06-04 01:08:50 [INFO] LR → Brier: 0.1234, ECE: 0.1372
2025-06-04 01:08:50 [INFO]   → RF: tuning
2025-06-04 01:08:50 [INFO]     RF → AUC: 0.8940, F1: 0.9373
2025-06-04 01:08:50 [INFO] RF → Brier: 0.0811, ECE: 0.0299
2025-06-04 01:08:50 [INFO]   → GB: tuning
2025-06-04 01:08:51 [INFO]     GB → AUC: 0.8739, F1: 0.9323
2025-06-04 01:08:51 [INFO] GB → Brier: 0.0893, ECE: 0.0643
2025-06-04 01:08:51 [INFO]   → SVM: tuning
2025-06-04 01:08:51 [INFO]     SVM → AUC: 0.8075, F1: 0.9383
2025-06-04 01:08:51 [INFO] SVM → Brier: 0.0973, ECE: 0.0781
2025-06-04 01:08:51 [INFO]   → MLP: tuning
2025-06-04 01:08:52 [INFO]     MLP → AUC: 0.8233, F1: 0.9330
2025-06-04 01:08:52 [INFO] MLP → Brier: 0.0984, ECE: 0.0818
2025-06-04 01:08:52 [INFO] 5) Training TabPFN (max_time=10s)
2025-06-04 01:08:52 [INFO] Loaded cached TabPFN
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151: FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be removed in 1.8.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151: FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be removed in 1.8.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/base.py:474: FutureWarning: `BaseEstimator._validate_data` is deprecated in 1.6 and will be removed in 1.7. Use `sklearn.utils.validation.validate_data` instead. This function becomes public and is part of the scikit-learn developer API.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151: FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be removed in 1.8.
  warnings.warn(
2025-06-04 01:11:17 [INFO]     TabPFN → AUC: 0.9168, F1: 0.9379
2025-06-04 01:11:17 [INFO] TabPFN → Brier: 0.0794, ECE: 0.0676
2025-06-04 01:11:17 [INFO] 6) RQ2: SHAP feature importances for each model (1 sample each)
2025-06-04 01:11:17 [INFO]   → Explaining LR with shap.Explainer
2025-06-04 01:11:17 [INFO]   → Explaining RF with shap.Explainer
2025-06-04 01:11:17 [INFO]   → Explaining GB with shap.Explainer
2025-06-04 01:11:17 [INFO]   → Explaining SVM with shap.Explainer
2025-06-04 01:11:20 [INFO]   → Explaining MLP with shap.Explainer
2025-06-04 01:11:20 [INFO]   → Explaining TabPFN with shap.Explainer
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151: FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be removed in 1.8.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151: FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be removed in 1.8.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/base.py:474: FutureWarning: `BaseEstimator._validate_data` is deprecated in 1.6 and will be removed in 1.7. Use `sklearn.utils.validation.validate_data` instead. This function becomes public and is part of the scikit-learn developer API.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151: FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be removed in 1.8.
  warnings.warn(
```

```
ExactExplainer explainer: 2it [02:13, 133.88s/it]
2025-06-04 01:13:34 [INFO]     Top SHAP feature for LR: completion_time (5.1799)
2025-06-04 01:13:34 [INFO]     Top SHAP feature for RF: day_premium (0.0867)
2025-06-04 01:13:34 [INFO]     Top SHAP feature for GB: attitude_Friendly (0.0000)
2025-06-04 01:13:34 [INFO]     Top SHAP feature for SVM: completion_time (0.0000)
2025-06-04 01:13:34 [INFO]     Top SHAP feature for MLP: realized_return (0.0000)
2025-06-04 01:13:34 [INFO]     Top SHAP feature for TabPFN: realized_return (0.0394)
2025-06-04 01:13:34 [INFO] 7) Decision Curve Analysis (DCA) for all models
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151:
FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be
removed in 1.8.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151:
FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be
removed in 1.8.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/base.py:474: FutureWarning:
`BaseEstimator._validate_data` is deprecated in 1.6 and will be removed in 1.7. Use
`sklearn.utils.validation.validate_data` instead. This function becomes public and is part
of the scikit-learn developer API.
  warnings.warn(
/Users/alexsmacbook/venv/lib/python3.13/site-packages/sklearn/utils/deprecation.py:151:
FutureWarning: 'force_all_finite' was renamed to 'ensure_all_finite' in 1.6 and will be
removed in 1.8.
  warnings.warn(
2025-06-04 01:16:18 [INFO]     DCA computed for models: ['attitude_Friendly',
'realized_return', 'friendly', 'completion_time', 'attitude_Hostile', 'spread',
'attitude_Neutral', 'historical_avg_completion_time', 'base_equity_value', 'day_premium',
'LR', 'RF', 'GB', 'SVM', 'MLP', 'TabPFN']
2025-06-04 01:16:18 [INFO]
=== FINAL TEST SET PERFORMANCE ===
2025-06-04 01:16:18 [INFO] LR     → AUC: 0.7171, F1: 0.9239
2025-06-04 01:16:18 [INFO] RF     → AUC: 0.8940, F1: 0.9373
2025-06-04 01:16:18 [INFO] GB     → AUC: 0.8739, F1: 0.9323
2025-06-04 01:16:18 [INFO] SVM    → AUC: 0.8075, F1: 0.9383
2025-06-04 01:16:18 [INFO] MLP    → AUC: 0.8233, F1: 0.9330
2025-06-04 01:16:18 [INFO] TabPFN → AUC: 0.9168, F1: 0.9379
2025-06-04 01:16:18 [INFO]
Total runtime: 7.5 minutes
2025-06-04 01:16:18 [INFO] === PIPELINE COMPLETE ===
```