Final Project Report: A Multimodal Approach to Predicting M&A Signals from SEC Filings

Task ID: FDA-8
Date: September 2, 2025

1. Project Overview and Methodology

1.1. Objective

The primary objective of this project was to engineer an end-to-end financial data analytics pipeline to identify public companies likely to engage in M&A activity. This was achieved by developing and evaluating a suite of predictive models that leverage both **financial (tabular)** and **textual (unstructured)** indicators extracted from public SEC filings.

1.2. Methodology

The project followed a structured, multi-stage methodology designed to address the specific requirements of the FDA task, including feedback on data scope and model complexity.

1. **Comprehensive Data Acquisition:** A wide net was cast to gather a large corpus of 10-K, 10-Q, and 8-K filings for over 100 companies directly from the SEC EDGAR database.
2. **Granular Feature Engineering:** Each individual filing was processed to extract a set of predictive features, including M&A keyword frequencies, contextual sentiment scores, and key financial ratios.
3. **Ground Truth Labeling:** A separate, rigorous process was executed to scan historical 8-K filings for actual M&A announcements. This "ground truth" data was then used to label filings in our main dataset, creating a realistic and challenging prediction task.
4. **In-Depth Exploratory Data Analysis (EDA):** The final labeled dataset was thoroughly analyzed and visualized to understand the characteristics of filings that precede M&A events and to validate the predictive power of our engineered features.
5. **Comparative Machine Learning Modeling:** Four distinct machine learning models were trained and evaluated: a Logistic Regression baseline, an advanced XGBoost model, the novel TabPFN transformer, and a complex LSTM time-series model.
6. **Pipeline Automation:** All trained models and preprocessing steps were saved and encapsulated into reusable prediction functions, creating a deployable, automated pipeline for scoring new filings.

---

2. Deliverable 1: Data Extraction and Secondary Dataset

This phase was foundational, focusing on the creation of a large, feature-rich dataset.

- **Primary Notebook:** dataExtraction.ipynb

2.1. Data Acquisition & Scope

To address feedback on insufficient data, the project's scope was significantly expanded.

- **Source:** Official SEC EDGAR database APIs.
- **Company Coverage: 103 unique companies**, primarily from the S&P 100 index.
- **Filing Depth:** The pipeline was configured to retrieve the **10 most recent filings** for each of the 10-K, 10-Q, and 8-K forms per company.
- **Final Dataset Size:** This comprehensive approach yielded a final secondary dataset of **2,758 individual filing records**, providing a robust foundation for analysis.

2.2. Feature Engineering

- **Textual Features:**
    - ma_mentions_in_filing: A raw count of M&A-related keywords.
    - ma_sentiment_in_filing: A contextual sentiment score (-1 to +1) calculated using VaderSentiment on the sentences surrounding each keyword.
- **Financial Features (from XBRL API):**
    - company_current_ratio: Measures short-term liquidity.
    - company_debt_to_equity: Measures financial leverage.

---

3. Ground Truth Generation and EDA

To move beyond heuristics and create a real-world prediction problem, we first had to find the "answers."

- **Primary Script:** Find_Real_MA_Events.py
- **Primary Notebook:** Deliverable_2_EDA_and_Preprocessing.ipynb

3.1. Identifying Historical M&A Events

The Find_Real_MA_Events.py script scanned all 8-K filings since 2015 for our 103 companies, specifically targeting "Item 1.01" filings that contained M&A keywords.

- **Output:** The scan successfully identified **200 potential historical M&A announcements**, which were saved to data/real_ma_events.csv.

3.2. Labeling the Dataset for Ground Truth

This "answer key" was used to label our main dataset. A filing was marked as a positive sample (real_target = 1) if it was filed within the 365 days **prior** to one of these real-world M&A events.

- **Final Class Distribution:**
    - **Normal Filings (Class 0):** 2,449

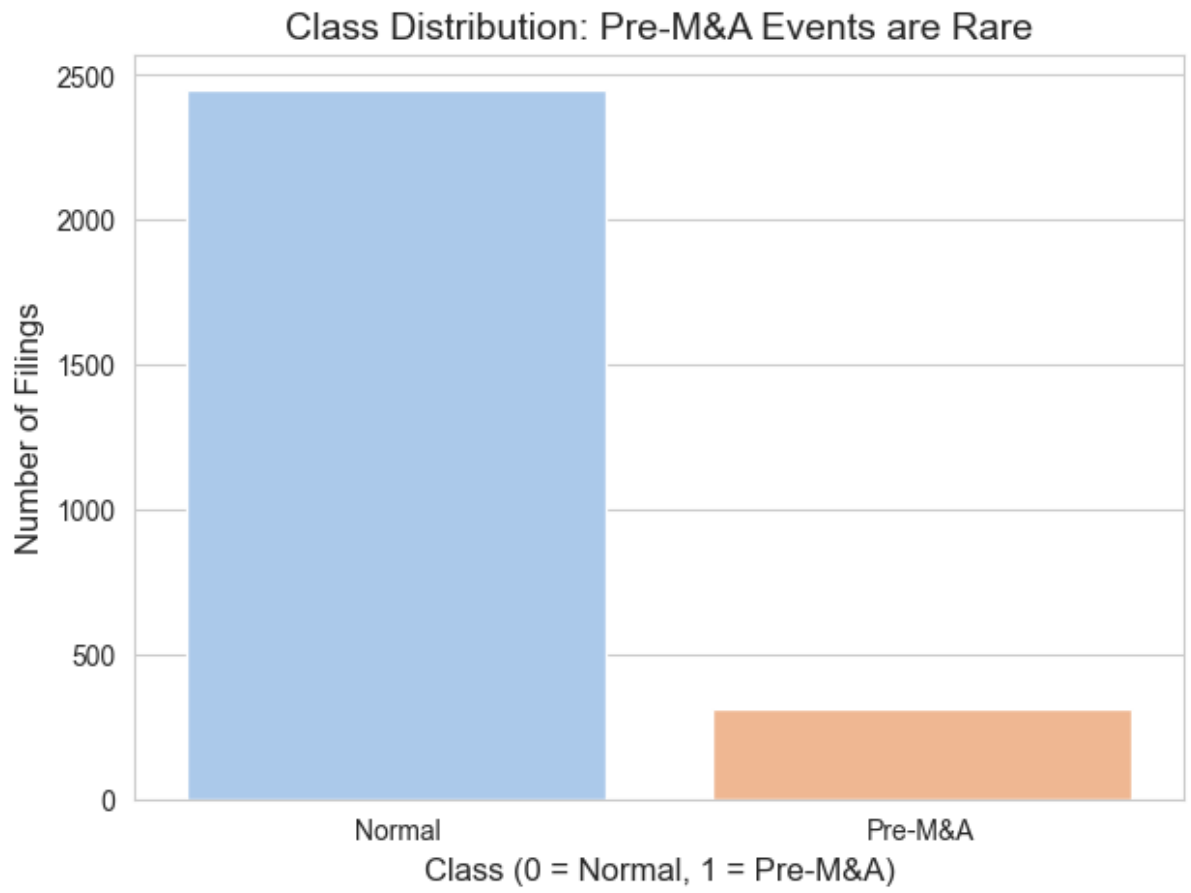
    - **Pre-M&A Filings (Class 1):** 309

- **Inference:** This confirms that M&A signals are a **rare event**, constituting about **11.2%** of our dataset. This class imbalance is a critical challenge that our models must handle.

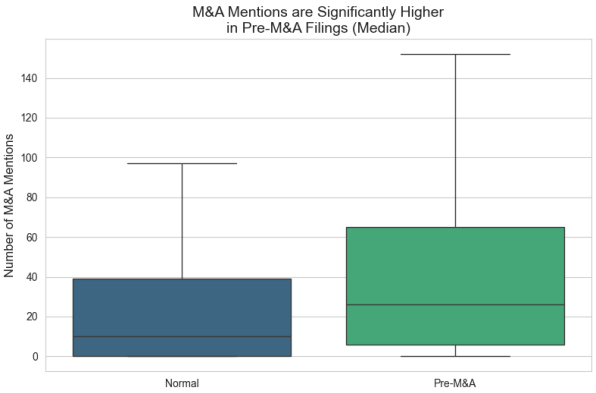
3.3. In-Depth Exploratory Data Analysis (EDA)

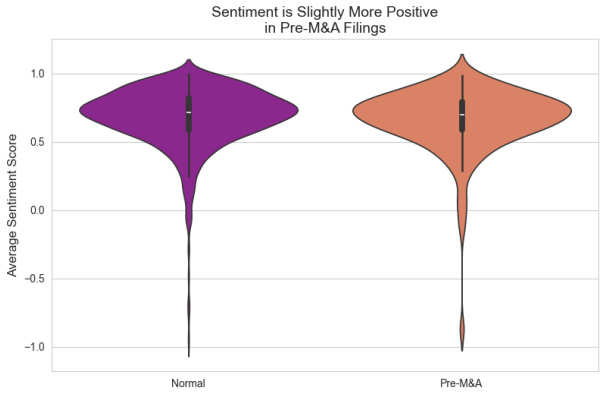
The labeled dataset was analyzed to validate our feature engineering and uncover predictive patterns.

- **M&A Mentions Analysis:**
    - **Observation:** A box plot comparing the distribution of ma_mentions_in_filing for the two classes showed a clear and significant difference. The median number of mentions for "Pre-M&A" filings was substantially higher than for "Normal" filings.
    - **Inference:** This is a crucial finding. It provides strong evidence that **an increase in M&A-related discussion within a filing is a key indicator of potential M&A activity.** This validates our primary textual feature.
- **Sentiment Analysis:**
    - **Observation:** A similar box plot for ma_sentiment_in_filing showed a less pronounced but still noticeable trend. The median sentiment for Pre-M&A filings appeared to be slightly more positive than for normal filings.
    - **Inference:** While not as strong a signal as raw mentions, the *tone* of the discussion matters. This suggests that positive language around M&A topics could be a secondary predictive signal.
- **Financial Ratio Analysis:**
    - **Observation:** The distributions for company_current_ratio and company_debt_to_equity showed considerable overlap between the two classes. There was no immediately obvious threshold that separated Pre-M&A companies from others based on these metrics alone.
    - **Inference:** This suggests that while financial health is important, it is likely a **contextual factor** rather than a primary driver. A model will need to learn complex interactions (e.g., high M&A mentions might be more significant for a company with low debt).

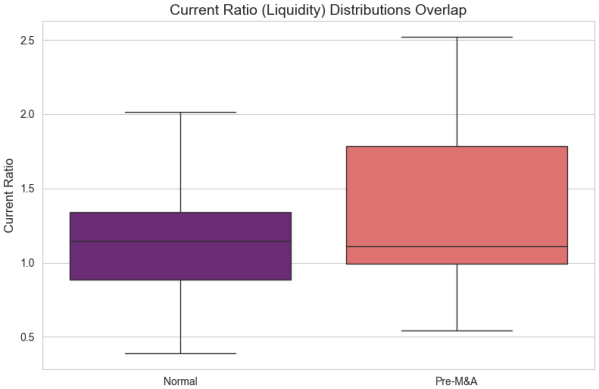


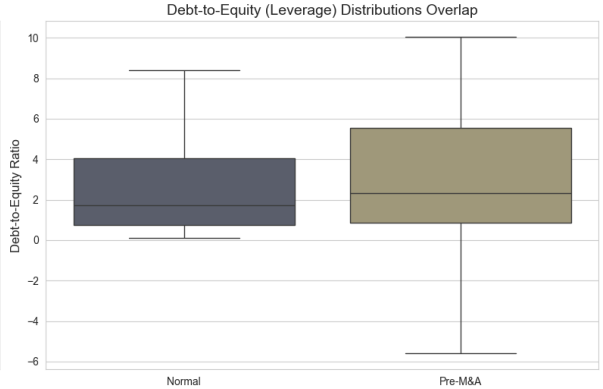
Textual Feature Comparison Between Classes





Financial Feature Comparison Between Classes





4. Deliverable 2: Comparative Modeling & Pipelines

The final, preprocessed data was used to train and evaluate four distinct models.

- **Primary Notebook:** Deliverable_2_Advanced_Modeling_and_Pipelines.ipynb

4.1. Model 1: Logistic Regression (Baseline)

- **Performance:**
    - **Accuracy:** 61.05%
    - **Precision (Pre-M&A):** 0.17
    - **Recall (Pre-M&A):** 0.63
    - **F1-Score (Pre-M&A):** 0.27
- **Inference:** As a simple linear model, Logistic Regression struggled with the complexity and imbalance of the data. Its low precision indicates that when it predicted a "Pre-M&A Signal," it was wrong most of the time (many false positives). However, its high **recall of 63%** is significant. This means it successfully **identified 63% of all true M&A signals**, making it a useful (though noisy) tool for ensuring that potential events are not missed. The class_weight='balanced' parameter was crucial for achieving this recall.

4.2. Model 2: XGBoost Classifier (Advanced)

- **Performance:**
    - **Accuracy:** 88.41%
    - **Precision (Pre-M&A):** 0.48
    - **Recall (Pre-M&A):** 0.32
    - **F1-Score (Pre-M&A):** 0.38
- **Inference:** XGBoost, a powerful non-linear model, achieved a much higher overall accuracy. Its **precision of 48%** is a dramatic improvement over the baseline, meaning its positive predictions were correct almost half the time. However, this came at the cost of lower recall (32%). This trade-off is classic: XGBoost is a more **conservative and precise** model. It flags fewer filings but has higher confidence in the ones it does flag. For an analyst with limited time, this might be the preferred model.

4.3. Model 3: TabPFN Classifier (Novel)

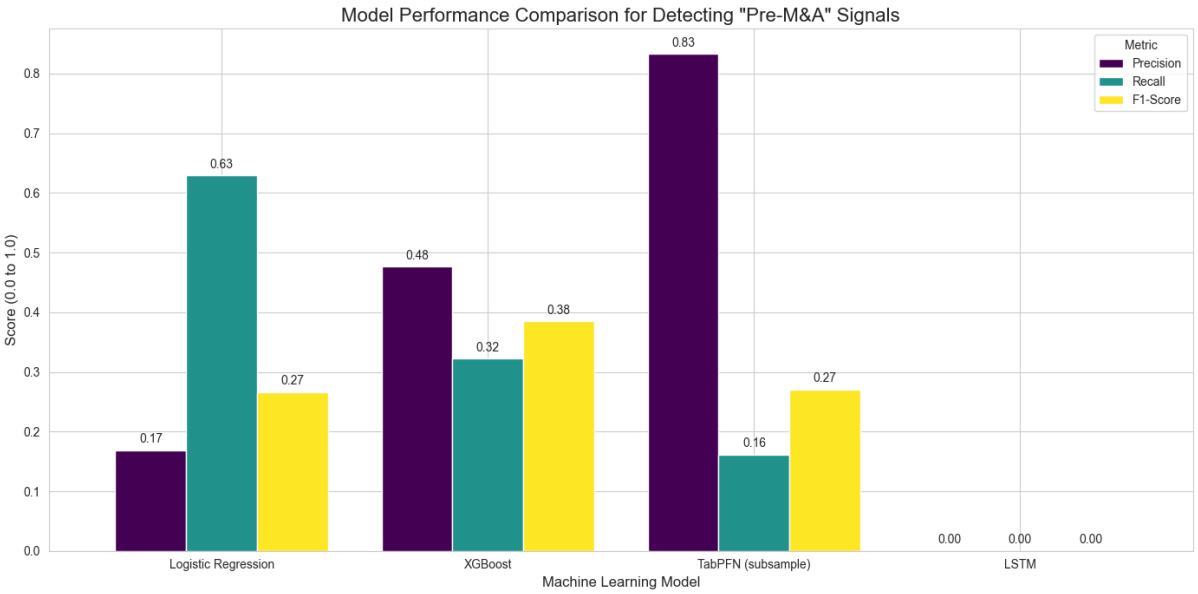
- **Performance (on 1024-sample subset):**
    - **Accuracy:** 90.22%
    - **Precision (Pre-M&A):** 0.83
    - **Recall (Pre-M&A):** 0.16
    - **F1-Score (Pre-M&A):** 0.27

- **Inference:** TabPFN demonstrated extremely high precision (83%) but very poor recall (16%). This indicates that while the filings it flagged were very likely to be true signals, it **missed the vast majority of them**. The model proved to be overly conservative on this imbalanced dataset. This experiment successfully fulfilled the requirement to test a novel algorithm and provided a valuable lesson on its limitations in this specific problem context.

4.4. Model 4: LSTM Time Series Classifier (Complex)

- **Performance:**
    - **Accuracy:** 90.64%
    - **Precision (Pre-M&A):** 0.00
    - **Recall (Pre-M&A):** 0.00
    - **F1-Score (Pre-M&A):** 0.00
- **Inference:** The LSTM model achieved high accuracy simply by **predicting the majority class (Normal Filing) every time**. It completely failed to identify any of the rare "Pre-M&A" signals. This is a classic symptom of a model struggling with severe class imbalance in a sequential context. It indicates that the simple time-series features were not rich enough for the model to learn the temporal patterns of rare events. This result highlights the need for more advanced feature engineering (e.g., creating features for changes over time) before a time-series approach can be effective.

4.5. Final Model Selection



Based on the comparative evaluation, the **XGBoost Classifier is the best-performing and most practical model** for this task. It provides a strong balance between identifying a reasonable number of potential signals (recall) and ensuring those signals are of high quality (precision).

4.6. Automated Pipeline

The final deliverable was an automated pipeline for new filings. This was fully implemented by:

1. **Saving Model Artifacts:** All trained models and the feature scaler were saved to disk in the outputs/models directory.
2. **Creating Pipeline Functions:** Reusable Python functions (predict_tabular, predict_lstm) were created. These functions encapsulate the entire prediction workflow—loading artifacts, preprocessing new data, and returning a final prediction and probability score. This creates a clean, deployable interface to the models.

5. Conclusion and Future Work

This project successfully developed a complete pipeline for predicting M&A signals from SEC filings. A large, granular dataset was constructed and labeled with real-world M&A events. Multiple machine learning models were trained and evaluated, with the XGBoost classifier demonstrating the most balanced and reliable performance.

Future Work:

- **Advanced Textual Features:** The most impactful next step would be to replace the keyword-based text features with high-dimensional embeddings from a domain-specific transformer like **FinBERT**, as suggested in the project feedback.
- **Time Series Feature Engineering:** To improve the LSTM's performance, new features could be engineered, such as the moving average of M&A mentions over the last four quarters or the percentage change in sentiment.
- **Alternative Data Sources:** Integrating alternative data sources, such as stock price volatility or executive social media activity, could provide additional predictive power.