GEORGE BROWN PRESS
JOURNALS, BOOKS & PROCEEDINGS

Journal of Computer, Signal, and System Research

Article

Index Weight Prediction and Capital Liquidity Analysis Based on Data Science

Minghao Chi [1,*]

[1] Global Markets Trading, Barclays Capital, New York, 10010, USA
* Correspondence: Minghao Chi, Global Markets Trading, Barclays Capital, New York, 10010, USA

Abstract: This study explores the intrinsic logical relationship between index construction methodologies and capital liquidity by leveraging advanced data science and computational technologies. We propose a multi-model hybrid framework for predicting index weight changes, incorporating diverse model sets to capture complex market dynamics. Key variables are systematically identified and screened through an integrated data platform and rigorous feature engineering, enabling the construction of a forward-looking index weight fluctuation pattern. This model serves as a foundational case study for examining the impact of index rebalancing behaviors on capital liquidity. Furthermore, we design early-warning mechanisms and clustering-based response strategies to anticipate liquidity risks, simulate stress scenarios, and develop a dynamic network transmission model that maps the propagation of market shocks. The results provide a comprehensive theoretical and practical reference for index management, risk mitigation, and the maintenance of financial market stability, offering valuable insights for both regulators and institutional investors.

Keywords: index weight prediction; capital liquidity; data science; multi-model integration; network conduction

1. Introduction

Indices serve as crucial price signals and guides for capital allocation, and with the continued deepening and sophistication of financial markets, changes in index composition and weightings have increasingly significant effects on market structure and liquidity dynamics. The construction of indices and the adjustment of their weights have become central topics in quantitative investment, portfolio management, and risk assessment research. Leveraging big data analytics and machine learning techniques, multi-agent prediction algorithms can integrate multi-source information to forecast weight adjustments more accurately, identify fund flow patterns, enhance market stability, and improve transparency. These predictive capabilities provide critical decision-support tools for both regulatory authorities and institutional investors, enabling more informed interventions and strategic planning.

For example, when Coinbase was successfully added to the S&P 500 index on May 16, 2025, its trading volume surged to 72 million shares on that single day-nearly ten times its average daily volume of 7 million shares. This dramatic spike illustrates the profound impact that index rebalancing events have on stock market liquidity, particularly affecting transactions dominated by passive investment funds. Such events highlight the urgent need for a predictive system capable of anticipating changes in index components, thereby facilitating proactive liquidity management and reducing systemic market risks. Developing such a system is not only a pressing practical necessity but also a critical step toward more resilient and adaptive financial markets.

Received: 07 September 2025
Revised: 19 September 2025
Accepted: 11 October 2025
Published: 14 October 2025

2. The Basis of the Relationship Between the Index Composition Mechanism and Capital Liquidity

2.1. Weight Distribution Mechanism

Index adjustments directly influence the redistribution of passive capital, as changes in component weights immediately affect the holdings of ETFs and index-tracking funds. In major global indices such as the MSCI or S&P 500, weight modifications often prompt funds to implement concentrated buy or sell orders within very short timeframes, leading to temporary surges in trading volume and short-term price volatility. Such operations place substantial demands on market liquidity. If the underlying assets exhibit low intrinsic liquidity, these adjustments can result in significant price deviations, elevated transaction costs, and amplified market shocks.

Incorporating liquidity considerations into the index construction and rebalancing process can therefore mitigate trading bottlenecks and preserve overall market stability. Beyond the magnitude of weight adjustments, additional factors-such as the frequency of index rebalancing, the timing of announcements, and the transparency of inclusion or exclusion criteria-also shape market participants' expectations regarding buying and selling behaviors. These expectations, in turn, can either exacerbate or dampen liquidity shocks arising from fund flows. Collectively, these phenomena underscore that index systems do not merely reflect market fundamentals but actively shape the flow of capital, making them a critical element in liquidity analysis and risk management frameworks.

2.2. Conduction Path of Weight Changes

Predicting the pressure of large-scale stock sell-offs through models of index weight adjustments is crucial for establishing effective liquidity early-warning mechanisms. Such predictive models typically draw upon historical weight adjustment data, observed market reactions, and recorded capital flows. For example, stocks experiencing significant weight increases often receive sustained net inflows of capital around the adjustment's effective date, whereas stocks with reduced weights or those being removed from the index frequently face net outflows over one or more trading sessions.

Machine learning techniques, including regression modeling and change-point detection, can identify key inflection points in these flows, enabling regulators and institutional investors to implement risk-sensitive, pre-emptive interventions. To improve predictive accuracy and responsiveness, short-term liquidity indicators-such as bid-ask spread fluctuations, transaction frequency, and order book depth-should be embedded in these models. By incorporating market microstructure data, the resulting dynamic feedback system can provide a comprehensive, agile, and real-time view of liquidity pressures at the trading level, enhancing both early-warning and risk mitigation capabilities.

2.3. Liquidity Indicators and Modeling Requirements

Swap transactions within index adjustments can further amplify liquidity pressures. For instance, Coinbase's inclusion in the S&P 500 on May 16, 2025, triggered a trading volume of 72 million shares, vastly exceeding its average daily volume of 7 million shares. This surge reflected the rapid participation of numerous passive investors, highlighting both the liquidity tolerance threshold of the market and the intensive demand generated by index rebalancing.

For smaller or less liquid companies, such concentrated inflows can create imbalances between supply and demand, increasing transaction costs and potentially causing abnormal price movements. To counteract such risks, buffering measures-such as multi-phase adjustment schedules and staggered implementation dates-can be incorporated into index rebalancing rules to enhance the market's self-regulatory capacity. Moreover, real-time monitoring of trading volumes for rebalanced assets enables regulators to detect emerging liquidity risks promptly and intervene as needed, reducing systemic structural vulnerabilities.

In sum, understanding the mechanisms of weight distribution, the conduction paths of changes, and the role of liquidity indicators is essential for designing robust index management frameworks. By integrating predictive modeling, real-time monitoring, and proactive buffering strategies, it is possible to mitigate liquidity shocks and promote sustainable market stability.

3. Construction of exponential weight prediction model based on data science methods

As the core analytical framework of this study, the index weight prediction model serves a dual purpose: it elucidates the internal dynamics of the index structure and provides essential input variables for subsequent analyses on how index rebalancing impacts market liquidity. By leveraging data-driven methods, this model captures both the temporal evolution of index components and the underlying mechanisms driving weight adjustments.

3.1. Data System Construction

Dynamic index weight changes represent the combined influence of market prices, pricing structures, and institutional arrangements. Building a comprehensive numerical framework is therefore fundamental to the prediction model. The required dataset can be categorized into three primary dimensions: market-level data, fundamental company data, and institutional data.

- **Market-level data** include indicators such as security prices, trading volumes, bid-ask spreads, order book depth, and capital flow metrics, which collectively capture market sentiment, liquidity conditions, and short-term trading pressures.
- **Fundamental company data** encompass firm size, revenue growth rates, valuation ratios, debt-paying ability, and other financial indicators that reflect the medium- to long-term investment value of constituent stocks within the index.
- **Institutional data** comprise index compilation rules, historical adjustments of constituent stocks, weight-determination methodologies, recalibration intervals, and the timing of index provider announcements. These data reveal the structural principles governing weight changes and encode the operational logic of index construction.

Given the heterogeneity of these datasets, preprocessing is critical. Cleaning, integration, and normalization must be performed based on a unified time-series index. Key tasks include filling missing values, smoothing boundary outliers, phase alignment, and time-frequency resampling, all of which ensure consistency, real-time performance, and data reliability. Additionally, constructing a feature matrix that captures both static attributes and dynamic changes is essential. This matrix forms the foundation for robust feature engineering, enabling the extraction of high-quality, predictive variables for subsequent modeling.

3.2. Feature Engineering and Variable Screening

Feature extraction constitutes a critical step in predicting index weights, as these derived features serve as key input variables for subsequent analyses, including liquidity assessment and risk modeling. The primary objective is to identify and construct structural indicators from the raw data that exhibit both high correlation and stability with respect to changes in index weights. In the context of time-series data, the sliding window technique is frequently employed to generate dynamic statistical features. Commonly utilized metrics include the n-day moving average (MA) to capture trend information, rolling standard deviation (Vol) to quantify volatility, the maximum-minimum ratio (Max/Min Ratio) to reflect relative price fluctuations, and price momentum (Momentum) to assess the speed and direction of market movements.

$$\mathrm{MA}_n(t) = \frac{1}{n}\sum_{i=0}^{n-1} P_{t-i}, \mathrm{Momentum}_n(t) = P_t - P_{t-n} \quad (1)$$

To further capture nonlinear structural changes in the market, additional features such as the abnormal trading volume ratio, volume-price correlation coefficient, net capital inflow rate, and short-term volatility index (SVI) are incorporated. These features enable the model to extract richer information about market microstructure dynamics and transient liquidity pressures. Following feature construction, redundancy control and importance verification are performed to ensure the quality and relevance of the input variables.

Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can be applied to these high-dimensional, complex variables, preserving the explanatory power of the principal components while reducing noise and multicollinearity. Alternatively, metrics such as information entropy, the Maximum Information Coefficient (MIC), and Pearson or Spearman correlation coefficients can be used to assess the impact of each feature on the target variable and to classify them into relevance tiers. Additionally, model-based feature importance evaluation methods allow for the identification of consistently influential factors, while mitigating the risk of overfitting during the feature selection process. Collectively, these procedures generate a high-quality and robust input matrix, laying a solid foundation for subsequent model training and predictive analysis.

3.3. Design of Exponential Weight Prediction Architecture for Multi-Model Integration

Given the nonlinear and dynamic nature of exponential index weights, a single predictive model is often insufficient to capture all potential trends and fluctuations. To address this limitation, a multi-model integration framework is employed to enhance both the stability and adaptability of predictions. The overall workflow encompasses data analysis, feature decomposition, multi-model training, and model fusion.

Within this framework, the underlying computational platform executes diverse model types in parallel to exploit their respective strengths. For instance, XGBoost is employed for processing structured features, LSTM networks capture temporal dependencies in the time-series data, and LightGBM handles high-dimensional feature interactions. Each model generates preliminary predictions independently, which are subsequently aggregated and optimized through a fusion layer-such as stacking, weighted averaging, or ensemble learning techniques-to produce the final estimated index weights (see Figure 1).

This multi-model architecture not only leverages complementary strengths across different algorithms but also mitigates the risk of model-specific biases and overfitting, resulting in a more robust and accurate predictive performance suitable for downstream liquidity analysis and risk management applications.

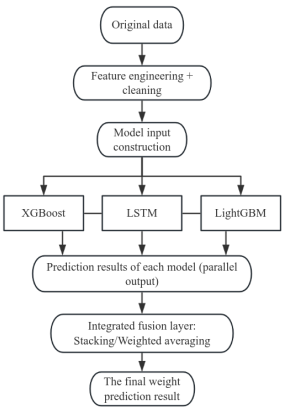

Figure 1. Flowchart of the exponential weight prediction architecture for multi-model integration.

This study employs the S&P 500 index as the validation dataset for the model, with data prior to January 1, 2020, designated as the training set and data from 2020 onwards

used as the test set to ensure rigorous out-of-sample evaluation. The LSTM model ingests multi-dimensional input data-including trading volume, volatility, turnover rate, and the direction of capital inflows and outflows-and outputs predicted changes in stock position weights over the subsequent five trading days. Since the S&P 500 undergoes quarterly rebalancing and the effective weights are determined based on the closing price of the reference day (e.g., the Q2 2025 rebalance effective on June 20 is determined by the June 13 close), directly predicting weights is both reasonable and practical. Prior to the reference day, constituent stocks and prices are uncertain, and machine learning models provide forward-looking estimates to address this uncertainty.

The model achieves a Mean Absolute Error (MAE) of 0.012 and an R^2 of 0.65 on the test set, indicating strong predictive performance, robust trend capture, and effective simulation capabilities. These predicted stock weights are then used as input parameters for constructing subsequent liquidity and market sentiment dissemination network models, thereby providing crucial support for downstream analysis.

3.4. Model Output Evaluation and Prediction Result Analysis Mechanism

Evaluating the model's output requires not only assessing numerical accuracy but also examining the underlying predictive mechanism and the robustness of the model structure. For exponential weight prediction, standard error metrics-such as Mean Square Error (MSE) and Mean Absolute Error (MAE)-quantify the numerical deviation of predictions. However, the direction of index weight changes has a more substantial impact on asset allocation decisions. Therefore, the Direction Accuracy (DA) metric should be incorporated to assess the correctness of predicted movement directions.

Furthermore, a Trend Consistency Index (TCI) can be constructed to evaluate the alignment of predicted and actual weight growth trends over continuous intervals. This index is defined as the ratio of predicted and actual weight changes that move in the same direction within a specified period, providing a quantitative measure of the model's ability to capture dynamic trend patterns in index weights. This comprehensive evaluation framework ensures that the model is not only accurate in magnitude but also reliable in capturing the temporal dynamics critical for portfolio and liquidity management.

$$TA = \frac{1}{T}\sum_{t=`}^{T\Sigma\left(\text{sign}\left(\hat{w}_i^{(t)}-\hat{w}_i^{t-1}\right)=sign\left(w_i^{(t)}-w_i^{(t-1)}\right)\right)} II \quad (2)$$

In this context, the indicative function is employed to determine whether the predicted and actual weight change directions are consistent. At the explanatory level, it is essential to quantify the causal pathways through which the integrated model influences the predicted values, thereby enhancing both transparency and interpretability. The SHAP (SHapley Additive exPlanations) algorithm is applied to assess the incremental contribution of each variable to the predictions, generating ranking graphs for both local and global feature importance. This facilitates the identification of the primary drivers of index weight changes.

For example, a significant proportion of the volume-price correlation coefficient in the global SHAP contribution highlights its substantial influence. By analyzing the underlying mechanisms, this variable can be interpreted as a mediator linking fund liquidity and the intensity of weight adjustments. Specifically, an increase in this coefficient suggests that investors' trading behaviors are becoming more synchronized, thereby amplifying market responses to the rebalancing of constituent stocks and resulting in more pronounced weight changes. These findings not only provide statistical validation but also elucidate the operational dynamics occurring during the rebalancing process.

Model output deviations are further monitored using an error residual analysis framework, particularly during periods of high market volatility or significant weight adjustments, to facilitate risk control. Implementing a rolling time window cross-validation strategy enhances temporal stability, allowing dynamic evaluation of model performance at each time point. The sample set is updated in real time to mitigate risks arising from

data drift, thereby ensuring the robustness and reliability of the model across multiple market cycles.

4. Analysis of the Liquidity Response Mechanism of Funds Based on Prediction Weights

4.1. Construction of Liquidity Early Warning Model

Building upon the predictive framework established in the previous section, this study investigates how redistributions of index weights-obtained through exponential weight forecasting-affect market capital flows. Forecasted index weights not only inform the asset allocation strategies of passive investment funds that track these indices but also provide a strategic "first-mover" advantage for anticipating market liquidity stress.

For example, on August 29, 2015, when Coinbase was included in the S&P 500 index, its stock price surged by nearly 20%, and trading volume spiked to more than ten times its normal level. This event illustrates the abnormal liquidity pressure caused by index rebalancing and highlights the critical role of prior prediction in mitigating market disruptions.

To address such liquidity risks, this study designs a liquidity early warning system capable of identifying potential disruptions in capital flows induced by index adjustments. The design process involves several steps:

1. **Feature Construction:** Key indicators include the magnitude and direction of predicted weight changes as well as their fluctuation trends, which collectively form the characteristic data for analysis.
2. **Integration of Market Activity Metrics:** Supplementary trading data-such as overall market capitalization, turnover rates, and trading activity levels-are incorporated to provide multidimensional inputs.
3. **Model Training:** Historical data are used to train classifiers or scoring models that quantify the risk levels for the subsequent period.
4. **Risk Output and Interpretation:** The system outputs graded risk levels-low, medium, or high-which can inform asset allocation decisions, trading strategies, regulatory monitoring, and other operational measures (see Figure 2).

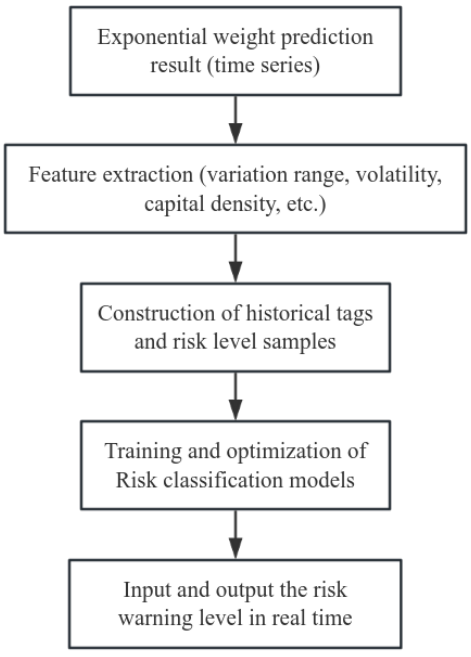

Figure 2. Flowchart of the liquidity early warning Model construction.

By systematically combining predictive weight information with market microstructure data, this early warning framework enables proactive identification and mitigation of liquidity stress, thereby enhancing market resilience and supporting informed decision-making for both investors and regulators.

4.2. Stress Scenario Path Simulation

The objective of stress scenario path simulation is to capture the nonlinear effects on market dynamics under extreme conditions induced by predicted changes in index weights. By constructing diverse shock scenarios, the transmission patterns resulting from abrupt fluctuations in price spreads, policy shifts, and large-scale capital inflows or outflows can be systematically analyzed. Changes in index weights reflect shifts in resource demand, which can, in turn, generate circulation pressures such as slippage, trading congestion, and settlement delays within the transaction execution system.

Beyond static price shocks, the simulation incorporates trading behaviors, including variations in pending order density, thinning of order book depth, and increased order cancellation activity. System parameters can be dynamically modeled, and feedback mechanisms introduced, allowing variables to adjust in real time according to market conditions-for example, subsequent redemptions triggered by price declines or amplified effects of passive ETF sales on the spot market.

To enhance the realism and robustness of the simulation, a wide range of market conditions is considered, encompassing periods of high volatility, low trading volume, and scheduled index adjustments. Multiple path sets are generated to systematically test system vulnerabilities under different stress levels. The simulation outputs include the distribution of liquidity risks, identification of extreme values, and delineation of potential system instability boundaries, providing actionable insights for risk mitigation and strategy formulation.

4.3. Liquidity Clustering and Dynamic Analysis

Fluctuations in index weights induce phased changes in market liquidity, often characterized by concentration and spatial-temporal transfer of liquidity. Unsupervised clustering models can classify the market into typical liquidity regimes-such as extremely high, moderate, restricted, and discrete liquidity states-allowing the detection and quantification of associated risk ranges. In this study, the market is segmented into four liquidity types, each reflecting distinct combinations of trading volume, price volatility, and transaction costs.

Cluster modeling relies on a multi-dimensional feature framework, encompassing price stability, trading behavior, order book depth, and transaction friction. This approach captures the complex interactions among market participants and allows for dynamic monitoring of liquidity transitions. Table 1 summarizes commonly used features for liquidity clustering, highlighting their relevance in identifying risk-prone periods and informing liquidity management strategies.

By combining stress scenario simulations with liquidity clustering, this framework provides a comprehensive and dynamic understanding of how index rebalancing propagates through market microstructure, supporting both preemptive risk management and the design of targeted intervention strategies.

Table 1. Explanation of the Liquidity Clustering Feature System and Dimensions.

Category	Feature example	Description
Transaction behavior	Rate of change in trading volume and density of trading value	Capture market participation and activity

Hanging order structure	The depth ratio of buy and sell orders and the tension index of the trading line	Reflect the liquidity carrying capacity
Cost indicator	The absolute value of the bid-ask spread and impact cost	Measure transaction frictions and the degree of market tension
Temporal fluctuation	5-minute volatility, VWAP deviation	Describe short-term price stability and trading pressure
Lag response	The weight predicts the transaction lag within the change window	Show the degree of delay in market response caused by predicted changes

Common clustering algorithms employed in liquidity analysis include K-means, Gaussian Mixture Models (GMM), and spectral clustering, among others. To capture the temporal evolution of liquidity states, the resulting state sequences can be modeled using Markov processes or Hidden Markov Models (HMMs), which provide a probabilistic framework for representing state transitions over time.

The clustering results indicate that the market can be classified into four distinct liquidity states:

1. High trading volume - low volatility,
2. Medium trading volume - medium volatility,
3. Low trading volume - high volatility,
4. Irregular or discrete type.

The observed proportions of these states in the sample are 28%, 36%, 22%, and 14%, respectively. By integrating these liquidity states with predicted index weight changes, a dynamic state response pattern can be established. This pattern provides valuable insights into how liquidity conditions evolve in response to index rebalancing and serves as a foundational input for subsequent network analysis and risk evaluation, enabling more precise monitoring and management of market liquidity.

4.4. Network Conduction Modeling and Stability Testing

The constituent stocks of an index form a potential liquidity network through industry affiliations, style correlations, and capital flow linkages. Deviations in the predicted weights can cause passive funds to propagate along these network paths, potentially triggering localized or even systemic liquidity shocks. By modeling this system as a weighted directed graph, critical transmission channels and high-risk nodes can be identified using advanced techniques such as graph neural networks (GNNs).

Building on this network representation, interference simulations are introduced to conduct robust stress tests of the liquidity network, encompassing scenarios such as abnormal node behavior, variations in edge weights, and structural changes in network topology. In parallel, the amplitude of fluctuations for each index component is calculated to evaluate the overall system resilience under different stress conditions (see Table 2). This framework provides a quantitative basis for assessing network vulnerabilities, identifying potential contagion pathways, and informing targeted risk mitigation strategies.

Table 2. Schematic Table of Stability Test Results under Network structure perturbation.

Disturbance scene	Changes in network connectivity(%)	Changes in PageRank at key nodes(%)	Average path length variation(%)	The probability of subgraph breakage	Estimated recovery time (days)
The weights were	+5.6	+18.3	+2.1	Low(5.4)	Medium(3)

collectively raised (top 10%)					
Weight reduction (single leading stock)	-12.4	+37.5	+6.9	Medium(22.1)	Slow(5)
The industry concentration has risen	+3.3	-6.5	-1.7	Extremely low(1.2)	Fast(1)
High-frequency capital disturbance simulation	-7.9	+14.8	+4.3	Medium(16.7)	Medium(2)
Random node removal (10%)	-20.5	-2.3	+9.1	High(41.8)	Slow(6)

Note: The recovery time estimation is based on the average period for the simulated capital flow to recover to 80% of the initial connectivity.

As shown in Table 2, structural changes exert a significant influence on the robustness and stability of the entire liquidity system. When critical nodes or major stock prices exhibit abnormal behavior, their effects are amplified, further increasing systemic risk. The probability of subgraph breakage is positively correlated with the network recovery time, indicating that once the market is disrupted, restoring the liquidity network may require substantial time. In contrast, changes in industry concentration tend to be gradual and relatively stable, with comparatively lower impact intensity. Although the simulation approaches employed vary, their results converge on a central insight: anticipated changes in index weights-particularly those resulting from recompositions-propagate indirectly through the financial flow network formed by constituent stocks. By treating predicted index weights as the initiating cause, this study establishes logical consistency and clear analytical focus across the different components of the model.

5. Conclusion

Guided by the concept of index structure and leveraging big data technologies, this study has developed a novel index weight prediction algorithm coupled with a capital liquidity response system, forming an integrated technical framework encompassing model design, feature engineering, early warning mechanisms, and network transmission analysis. This framework not only provides a theoretical basis for analyzing and controlling the evolution of complex financial system structures but also offers practical technical support for market participants and regulators. By linking predictive modeling with liquidity network analysis, the study delivers actionable insights into how index adjustments propagate through financial markets, enhancing both market stability and risk management capabilities.

References

1. W. S. Chen, and Y. K. Du, "Using neural networks and data mining techniques for the financial distress prediction model," *Expert systems with applications*, vol. 36, no. 2, pp. 4075-4086, 2009.
2. A. Riabykh, I. Suleimanov, I. Nagovitcyn, D. Surzhko, M. Konovalikhin, and O. Koltsova, "Entropy-based text feature engineering approach for forecasting financial liquidity changes," *EPJ Data Science*, vol. 14, no. 1, p. 17, 2025. doi: 10.1140/epjds/s13688-025-00535-z.
3. Y. Xiao, X. Zhao, and Y. Wu, "Research and Optimazation on Evaluation Index Prediction Using Grey Neural Network and Big Data," In *2021 IEEE International Conference on Electronic Technology, Communication and Information (ICETCI)*, August, 2021, pp. 233-237. doi: 10.1109/icetci53161.2021.9563267.

4. H. J. Im, S. Selvam, and K. J. Tan, "Effect of stock liquidity on the economic value of patents: Evidence from US patent data," *International Review of Financial Analysis*, vol. 94, p. 103314, 2024.
5. C. Mahony, M. Manning, and G. Wong, "Weighting Justice Reform Costs and Benefits Using Machine Learning and Modern Data Science," *World Bank*, 2023. doi: 10.1596/1813-9450-10449.