Proceedings of FSE 570 Data Science Capstone
School of Computing and Augmented Intelligence
Arizona State University
January - May, 2025, Tempe, USA

WHAT MAKES AN IPO SUCCESSFUL

Deepakshi Mathur
Arizona State University
Tempe, Arizona 85288
Email: dmathur4@asu.edu

Gunner Tillemans
Arizona State University
Tempe, Arizona 85288
Email: ghtillem@asu.edu

Julian Lin
Arizona State University
Tempe, Arizona 85288
Email: ylin389@asu.edu

Ninad Shah
Arizona State University
Tempe, Arizona 85288
Email: nshah103@asu.edu

Shradda Pandey
Arizona State University
Tempe, Arizona 85288
Email: spande81@asu.edu

ABSTRACT

This study presents a machine learning-driven analysis of key factors influencing the short-term and long-run success of Initial Public Offerings (IPOs). Utilizing structured disclosure data from the SEC EDGAR database, macroeconomic indicators, and market trends, a deep neural network and LightNGB model was developed to predict IPO success. Key financial ratios, growth metrics, and market conditions were engineered as predictive features. The Neural Network model achieved a test accuracy of approximately 69%, and The LightNGB model achieved a test accuracy of approximately 74%, with SHAP analysis identifying CPI, SnP500 200 day SMA, interest rates, Company type, and assets as critical drivers. These findings offer actionable insights for investors and financial analysts in evaluating upcoming IPOs. Future extensions include sector-specific modeling and integration of sentiment analysis.

NOMENCLATURE

IPO Initial Public Offering, when a private company offers shares to the public for the first time.
SEC Securities and Exchange Commission, the U.S. regulatory body overseeing securities markets.
EDGAR Electronic Data Gathering, Analysis, and Retrieval system for SEC filings.
FRED Federal Reserve Economic Data, a source of U.S. economic indicators.
S&P 500 Standard & Poor's 500 Index, representing U.S. market conditions.
CIK Central Index Key used to uniquely identify SEC filers.
SMA Simple Moving Average, a technical indicator used for stock trends.
SHAP SHapley Additive exPlanations, method for model interpretability.

INTRODUCTION

Initial Public Offerings (IPOs) are critical milestones that enable private firms to access public capital and expand their market presence. However, many newly listed companies underperform within months, making IPO success difficult to predict. For investors, underwriters, and policymakers, understanding the drivers of IPO outcomes is crucial to managing risk and capital allocation.

Traditionally, IPO evaluation has relied on isolated financial metrics or qualitative judgment, often missing the complex interactions among firm fundamentals, market sentiment, and macroeconomic conditions. This project addresses that gap by developing a machine learning-based model that predicts IPO success using structured financial disclosures, economic indicators, and historical market data.

Acknowledging that Predicting IPO performance isn't your standard classification task, but rather a complex, multi-dimensional challenge involving noisy data, shifting market conditions, and delayed success indicators. our goal with this project is to add a little predictability in the broader understanding of value creation in IPOs given it being an overwhelmingly unpredictable domain.

DATA SOURCES AND ANALYSIS

To develop a predictive model for IPO success, multiple high-quality datasets were integrated. Financial statement data were extracted from the SEC EDGAR database, covering IPO filings and quarterly reports from 2009 to 2024. We used separate SEC repositories to connect the CIK to Tickers. Historical stock prices were retrieved from EODHD Finance, enabling the calculation of performance metrics at six months and three years post-IPO. Macroeconomic indicators, including inflation rates and Federal Funds rates, were obtained from the Federal Reserve Economic Data (FRED) platform. Market sentiment proxies were derived using Yahoo Finance from S&P 500 SMA trends. Lastly we were granted access to OpenCorperates and extracted companies date of incorporation.

Data Collection

To build a robust IPO success prediction model, diverse and high-quality data spanning 2009 to 2024 were collected from authoritative sources across financial, market, and macroeconomic domains.

Company-level financial data were sourced from the SEC EDGAR database, including S-1 registration statements, 10-K annual reports, and 10-Q quarterly reports. Each firm's data were identified using Central Index Keys (CIKs), which were mapped to IPO dates and corresponding stock tickers to enable temporal alignment and integration.

Historical stock prices were retrieved from EODHD Finance at three key intervals: the IPO date, six months post-IPO, and three years post-IPO. These prices were merged with financial disclosures using the CIK–ticker mapping.

Macroeconomic context was captured using data from the Federal Reserve Economic Data (FRED) platform, including prevailing inflation and Federal Funds rates at each IPO date. To assess market sentiment, S&P 500 Index trends were analyzed based on the index's position relative to its 30-day, 50-day, and 200-day moving averages. Lastly we extracted the companies age of incorporation from OpenCorprates to calculate the age of the company when the IPO opened.

This unified dataset, integrating financial, market, and macroeconomic indicators, provided a comprehensive foundation for downstream prepossessing and modeling.

Data Preprocessing and Feature Engineering

Data Preprocessing Given the heterogeneous nature of the collected datasets, careful and rigorous preprocessing was essential to ensure consistency, reliability, and analytical readiness for modeling.

Raw SEC EDGAR filings, initially stored across multiple relational tables ('num' for numeric financial entries, 'tag' for financial metric definitions, and 'sub' for submission metadata), were systematically joined to reconstruct complete historical financial statements for each IPO candidate. A comprehensive Central Index Key (CIK) to ticker mapping was developed using authoritative SEC metadata sources, enabling seamless integration between financial disclosures and corresponding stock price records retrieved from EODHD Finance.

Temporal filtering was employed to focus exclusively on financial data preceding each company's IPO date, ensuring that predictive features represented only historical information available before the public offering. Adjusted closing stock prices from EODHD Finance were precisely aligned relative to each company's IPO date to establish consistent evaluation points (initial listing, six months post-IPO, and three years post-IPO).

Macroeconomic indicators such as inflation rates and Federal Funds rates from the Federal Reserve Economic Data (FRED) platform were mapped to IPO listings based on the month and year of the public offering. Additionally, market sentiment indicators were created by comparing the S&P 500 Index price at the IPO date to its 30-day, 50-day, and 200-day simple moving averages (SMAs).

The dataset we compiled comprised structured data for 5,057 Initial Public Offerings (IPOs). It integrated firm-level financial fundamentals, industry-specific metadata, and broader macroeconomic indicators recorded around the IPO date. Data was sourced from various databases and merged into a unified dataset, after excluding columns with 50% or more missing values resulted in 121 initial features for model training, as rigorous data cleaning was performed to mitigate potential data leakage and improve predictive reliability.

All numerical features underwent standardization to achieve zero mean and unit variance. Missing or anomalous values arising from incomplete or inconsistent data submissions were systematically addressed through appropriate imputation methods or exclusions based on their frequency and significance.

Specifically, the following preprocessing steps were executed:

1. **Removal of Leakage Features:** Columns containing direct future indicators or post-event metrics, such as future stock prices or subsequent market indicators (e.g., `six_month_price`, `three_year_price`, and moving averages calculated post-IPO), were explicitly removed to prevent data leakage.
2. **Conversion of Categorical and Boolean Features:** All cat-

egorical features initially stored as object types were coerced into numeric format where possible, while boolean columns were explicitly converted to integers (0 and 1).

3. **Missing Data Imputation:** Numerical features underwent median imputation to manage missing values, defined as:

$$x_{\text{imputed}} = \text{median}(X)$$

Categorical features had missing values imputed using a constant placeholder value of -1, facilitating their handling by LightGBM.

4. **Outlier Clipping:** To address potential distortion by extreme outliers, numerical values were clipped using the interquartile range (IQR) method:

$$x_{\text{clipped}} = \min\left(\max(x, Q_1 - 1.5 \times \text{IQR}), Q_3 + 1.5 \times \text{IQR}\right),$$

where Q_1 and Q_3 represent the 25th and 75th percentiles, respectively, and:

$$\text{IQR} = Q_3 - Q_1.$$

Post-cleaning, the final dataset comprised 103 predictive features.

This comprehensive preprocessing pipeline transformed the raw, disparate datasets into a unified, high-integrity dataset optimized for subsequent feature engineering and predictive modeling of IPO success.

Feature Engineering Feature engineering focused on extracting predictive signals from the pre-processed data sources:

Pre-IPO Financial Features: Daily rate of change (ROC), mean, and median values were computed for key financial tags (e.g., assets, revenue) over multiple pre-IPO windows (1Y, 2Y, 3Y, full history). These metrics were derived per quarter and reshaped into wide-format features (e.g., `Q4_Assets_2Y_ROC`). Zero values were treated as missing where appropriate.

The resulting metrics were then pivoted into a wide format, creating features such as *`Qqtrs_tag`* (e.g., `Q4_Assets_2Y_ROC`, `Q1_Revenue_1Y_mean`), yielding one row per CIK summarizing its pre-IPO financial trajectory across various metrics and timeframes. Zero values in financial tags were treated as missing (replaced with `NaN`) during these calculations where appropriate for the specific metric (e.g., before melting for ROC/mean/median calculations).

Market Trend Features: Prevailing market sentiment at the time of the IPO was proxied by comparing the S&P 500 index value on the IPO date to its 30-day, 50-day, and 200-day simple moving averages (SMAs). Ratios or difference indicators were derived to label if market was above or below SMA.

Macroeconomic Features: The collected FRED inflation and Federal Funds rates corresponding to the IPO date were directly included as features representing the macroeconomic climate at the time of IPO opening.

Success Label Definition: The primary outcome variable, representing IPO success, was defined based on post-IPO stock price performance. An IPO was classified as a 'success' if it achieved a stock return of at least 6% within six months and/or at least 40% growth within three years, relative to the initial offer price (see Methodology section for details). This success label served as the target variable for the predictive models.

Final Dataset Assembly and Standardization: All engineered features derived from pre-IPO financials, market sentiment, and macroeconomic indicators were consolidated into a final analytical dataset. Missing values resulting from unreported financial tags or merge constraints were handled through imputation or exclusion based on context. All numeric features were standardized to zero mean and unit variance ($\mu = 0, \sigma = 1$) to ensure consistency during model training. This end-to-end pipeline transformed raw, multi-source data into a high-integrity dataset tailored for predictive modeling of IPO outcomes.

METHODOLOGY AND APPLICATION

Methods

Target Variable Definition We defined two binary target variables, each indicating IPO success at different intervals post-offering:

Short-Term Success (flag_6m_avg): Defined as a positive stock performance (above average market returns) at six months post-IPO.
Long-Term Success (flag_3y_avg): Defined similarly, but evaluated at three years post-IPO.

Each model was developed independently, tailored specifically to the respective timeframe.

Modeling Framework: Gradient Boosting with LightGBM We employed LightGBM, a gradient boosting framework optimized for structured tabular data. Gradient boosting builds sequential ensembles of decision trees, with each new tree constructed to minimize residual errors from previous iterations. Formally, the ensemble prediction $\hat{y}_i$ for observation i is:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathscr{F}$$

where f_k represents individual decision trees in the ensemble, and $\mathscr{F}$ denotes the function space defined by the tree structures.

The objective function optimized by LightGBM for binary classification is the binary logistic loss:

$$\mathscr{L}(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right],$$

where y_i and $\hat{y}_i$ are the actual and predicted probabilities of success, respectively.

Hyperparameter Optimization using Optuna To identify optimal hyperparameter combinations for robust predictive performance, we utilized the Bayesian optimization framework provided by Optuna. The hyperparameters tuned via Optuna included:

Boosting method: $\{\text{gbdt}, \text{dart}, \text{goss}\}$
Number of estimators (trees): $n_{\text{estimators}} \in [200, 2500]$
Learning rate: $\{0.005, 0.01, 0.05, 0.1\}$
Number of leaves per tree: $num\_leaves \in [10, 150]$
Maximum tree depth: $max\_depth \in [3, 15]$
L1 regularization coefficient: $\alpha \in [10^{-3}, 10^{1}]$
L2 regularization coefficient: $\lambda \in [10^{-3}, 10^{1}]$
Subsample ratio of data instances: $subsample \in [0.6, 1.0]$
Subsample ratio of features per tree: $colsample\_bytree \in [0.5, 1.0]$
Minimum data points in leaf nodes: $min\_child\_samples \in [5, 100]$
Minimum sum of instance weights per child: $min\_child\_weight \in [10^{-3}, 10^{1}]$

Optuna evaluated each trial configuration using **Repeated Stratified K-Fold Cross-Validation** (5 folds, repeated twice), ensuring robust assessment and reduced variance due to data splits. Class imbalance was addressed explicitly through the parameter:

$$\text{scale\_pos\_weight} = \frac{N_{\text{negative}}}{N_{\text{positive}}},$$

thus weighting positive samples proportionally higher to address imbalance.

Early stopping was employed (50 rounds) to mitigate overfitting. The primary evaluation metric during optimization was the Area Under the ROC Curve (AUC):

$$\text{AUC} = \int_0^1 \text{TPR}(FPR^{-1}(x))\,dx$$

Neural Network-Based IPO Success Prediction In addition to tree-based models, we developed a feedforward neural network classifier to predict IPO success at two future intervals: six months and three years post-offering. The models were implemented using TensorFlow and optimized via Optuna, a state-of-the-art Bayesian hyperparameter tuning framework.

Data Preparation and Preprocessing The same cleaned dataset used for the LightGBM model was reused for the neural network. After excluding columns flagged as data leakage risks, features were separated into numerical and categorical types. The preprocessing pipeline included:

Numerical Features: Imputed using the median, standardized via Z-score normalization:

$$x_{\text{scaled}} = \frac{x-\mu}{\sigma}$$

where μ and σ are the mean and standard deviation of the feature, respectively. An optional Principal Component Analysis (PCA) transformation was also applied, with the number of components treated as a tunable parameter.
.

Categorical Features: Imputed with the constant value `missing`, then one-hot encoded.

Preprocessing was handled through a `ColumnTransformer`, and the resulting transformed matrices were passed to the neural network as input features.

Network Architecture and Optimization The neural network followed a simple yet flexible multi-layer perceptron (MLP) structure with the following components:

$$\begin{aligned}
\mathbf{h}_1 &= \text{Dropout}_{p_1}\big(\text{BatchNorm}\big(\text{ReLU}(\mathbf{W}_1\mathbf{x}+\mathbf{b}_1)\big)\big)\\
\mathbf{h}_2 &= \text{Dropout}_{p_2}\big(\text{BatchNorm}\big(\text{ReLU}(\mathbf{W}_2\mathbf{h}_1+\mathbf{b}_2)\big)\big)\\
\hat{y} &= \sigma(\mathbf{W}_3\mathbf{h}_2+\mathbf{b}_3)
\end{aligned}$$

The network was trained using the Adam optimizer with a tunable learning rate and L2 weight regularization (weight decay):

$$\mathscr{L}(\mathbf{y},\hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] + \lambda \sum_{l} \|\mathbf{W}_l\|_2^2$$

Hyperparameter Tuning via Optuna A nested cross-validation scheme was used where hyperparameter tuning was done over the training set using `StratifiedKFold` with 2 folds. Optuna explored the following hyperparameter space:

Number of PCA components (`pca_n`)
Hidden layer units: `units1` $\in [32,256]$, `units2` $\in [16,128]$
Dropout rates: `drop1`,`drop2` $\in [0.1,0.5]$
Learning rate: `lr` $\in [10^{-5},10^{-2}]$
L2 regularization: `wd` $\in [10^{-6},10^{-2}]$

Each trial was trained for up to 100 epochs, with early stopping based on validation AUC (patience = 10). The Optuna objective maximized mean AUC across folds.

Final Model Training and Evaluation Once optimal hyperparameters were identified, the model was retrained on the entire training and validation set (with 20% reserved internally for monitoring) and then evaluated on the held-out test set. The final performance metric was the Area Under the ROC Curve (AUC), computed as:

$$\text{AUC} = \int_0^1 \text{TPR}(FPR^{-1}(x))\,dx$$

Final models and preprocessing pipelines were saved using the Keras and `joblib` serialization formats, respectively. This architecture facilitated both reproducibility and deployment for further analysis.

Final Model Training and Evaluation The best hyperparameter configuration from Optuna was utilized for training a final LightGBM model on the entire training dataset. To further prevent overfitting, 15% of the training data was withheld as a validation subset for early stopping.

The final trained models were evaluated against the reserved test dataset (30% of the original dataset), providing an unbiased estimate of their predictive performance on new, unseen data.

Feature Importance Assessment To interpret model predictions and identify critical predictors of IPO success, feature importance scores were calculated. LightGBM computes feature importance as the number of times a feature is utilized in tree splits (split frequency) and as the gain in information from splits (gain importance):

$$\text{Gain}(x_j) = \sum_{\text{all splits on } x_j} \text{Information\_Gain}$$

We reported and analyzed the top predictive features identified by this approach, providing insights into economic indicators and financial attributes most predictive of IPO success.

Model Evaluation and Simulation

Two predictive modeling approaches were evaluated: Light Gradient Boosting Machines (LightGBM) and Neural Networks (NN). Both models targeted IPO success at six-month and three-year intervals post-IPO.

The LightGBM models utilized ColumnTransformer pipelines for feature preprocessing, including median imputation and outlier clipping for numerical data, and categorical encoding with constant imputation. Stratified train-test splits ensured unbiased evaluation, and hyperparameters were optimized using the Optuna framework with repeated stratified K-fold cross-validation. Early stopping criteria were employed to prevent overfitting.

The NN models employed dimensionality reduction via Principal Component Analysis (PCA), followed by dense neural network architectures. These models underwent extensive hyperparameter tuning through Optuna, optimizing parameters such as PCA components, hidden layer sizes, dropout rates, learning rates, and weight decay values.

To robustly evaluate predictive efficacy, we conducted an investment simulation comparing three distinct agents: the Random agent, the LightGBM agent, and the NN agent. Each agent started with an initial capital of $100,000 and could invest up to $200 per IPO opportunity. Investment decisions were structured as follows:

Random Agent: Stochastically invests based on predefined probabilities (30% for six-month success, 20% for three-year success).
LightGBM and NN Agents: Invest $100 if success is predicted for either the six-month or three-year interval. Invest $200 (split equally) if success is predicted for both intervals.

Investment returns were calculated based on the actual stock prices at the six-month and three-year marks post-IPO. The simulation tracked cumulative profits and final capital to provide a clear economic assessment of each predictive approach.

RESULTS

Neural Network

Neural network (NN) models were trained separately to predict IPO success at two distinct intervals: six months and three years post-IPO. Both models underwent extensive hyperparameter optimization using Optuna, employing stratified cross-validation to select optimal parameters based on the area under the receiver operating characteristic curve (AUC).

For predicting six-month IPO success (`flag_6m_avg`), the best-performing model achieved a cross-validation AUC of 0.6712. After final training and evaluation on the held-out test set, this model achieved a test AUC of 0.6823. The optimal hyperparameters included 84 principal component analysis (PCA) components, two hidden layers with 96 and 48 neurons respectively, dropout rates of approximately 0.48 and 0.23, a learning rate of 0.0087, and a weight decay of 0.00027.

For predicting three-year IPO success (`flag_3y_avg`), the best model attained a higher cross-validation AUC of 0.7186. Final evaluation on the test dataset yielded an AUC of 0.6997. This model utilized 33 PCA components, two hidden layers containing 256 and 96 neurons respectively, dropout rates around 0.34 and 0.16, a significantly lower learning rate of 0.0002, and minimal weight decay of approximately 0.000015.

TABLE 1. Neural Network Classification Performance

Target	Accuracy	Precision	Recall	F1 Score	AUC
3-Year (`flag_3y_avg`)	0.5949	0.2905	0.7671	0.4214	0.7056
6-Month (`flag_6m_avg`)	0.6877	0.7500	0.0126	0.0247	0.6777

Model Results

The LightGBM classifier outperformed the neural network across both prediction horizons. For the three-year success label, LightGBM achieved an accuracy of 74.3% and an AUC of 0.7714, compared to the NN's accuracy of 59.5% and an AUC of 0.7056. At the six-month horizon, LightGBM reached an accuracy of 68.3% and an AUC of 0.7509, while the NN managed a slightly higher accuracy of 68.8% but had a significantly lower AUC of 0.6777, largely due to a very low recall for successful IPOs.

TABLE 2. LightGBM Classification Performance

Target	Accuracy	Precision	Recall	F1 Score	AUC
3-Year (`flag_3y_avg`)	0.7431	0.3855	0.5651	0.4583	0.7714
6-Month (`flag_6m_avg`)	0.6825	0.4969	0.6757	0.5727	0.7509

Feature Importance Analysis for Predictive Models

Combined Analysis of 3-Year and 6-Month Average Flag Models: The feature importance analysis for the `flag_3y_avg` and `flag_6m_avg` models, each evaluated on 1518 test rows, reveals significant overlap in key predictors, with `start_price`, `sp500_sma50_at_start`, `sp500_close_at_start`, `sp500_sma200_at_start`, `fed_funds_rate_at_start`, `inflation_rate_at_start`, and `sic_description_Blank Checks` ranking highly in both. These shared features underscore the critical influence of initial market conditions, macroeconomic factors, and industry-specific dynamics on flag outcomes over both short and long terms. The `flag_3y_avg` model emphasizes long-term financial stability through metrics like `Q0_CommonStockValue_2Y_mean`, `Q0_AssetsCurrent_2Y_mean`, and `Q0_Assets_2Y_ROC`, while the `flag_6m_avg` model prioritizes liquidity with features such as `Q0_LiabilitiesCurrent_2Y_median` and `Q0_CashCashEquivalentsRestrictedCashAnd RestrictedCashEquivalents_2Y_mean`. The prominence of shorter-term market indicators like `sp500_sma50_at_start` in the 6-month model, compared to a balanced reliance on both `sp500_sma50_at_start` and `sp500_sma200_at_start` in the 3-year model, highlights a greater sensitivity to recent market trends in shorter horizons. These findings suggest that while market and economic drivers are universally impactful, the time horizon shapes the relative importance of financial stability versus liquidity, offering insights for tailoring predictive strategies to specific forecasting periods.

Simulation Results

The simulation compared four strategies Random selection, LightGBM (LGBM), Neural Network (NN), and an Ensemble—starting from $100,000 and investing $100 per signal across 1,363 opportunities with a 0.60 prediction threshold.

LightGBM achieved the highest final capital of $161,116.00 (a $61,116.00 profit), outperforming the Ensemble ($140,652.26 final, $40,652.26 profit), the NN model ($139,850.54 final, $39,850.54 profit), and Random ($106,334.06 final, $6,334.06 profit). These results underscore LightGBM's superior ability to capture predictive signal in this IPO investment simulation.

Additional performance metrics highlight the distinct strengths of each model. The LightGBM model achieved the highest win rate (75.79%) and a robust win/loss ratio of 3.13. The Ensemble strategy, however, demonstrated the best risk management—posting the lowest maximum drawdown of $168.49—and delivered the highest average win per trade at $261.73. The Neural Network stood out on a risk-adjusted basis with the highest Sharpe ratio of 8.72, while the Random strategy, performed poorly risking the most capital and resulting in the lowest avg returns.

TABLE 3. Simulation Performance Metrics

Metric	Random	LGBM	Neural Network	Ensemble
Win Rate (%)	45.23	75.79	58.52	72.03
Total Trades	577	413	364	236
Average Win ($)	86.70	210.63	233.77	261.73
Average Loss ($)	-51.56	-48.11	-65.84	-58.22
Win/Loss Ratio	0.83	3.13	1.41	2.58
Max Drawdown ($)	3954.83	245.78	663.83	168.49
Avg Investment ($)	111.41	134.62	101.37	102.11
Sharpe Ratio	3.50	2.83	8.72	7.09

VISUALIZATION RESULTS

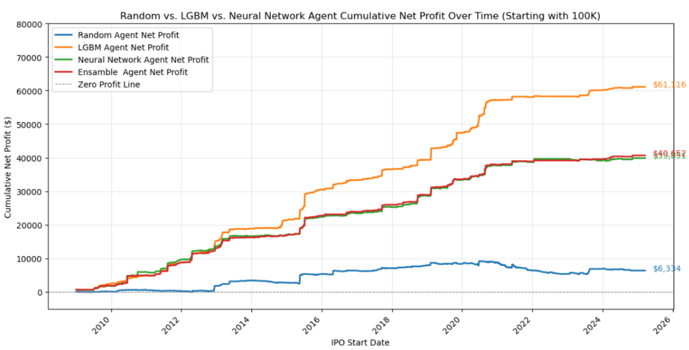

FIGURE 1. Cumulative net profit of Random, LGBM, Neural Network, and Ensemble trading agents (2009–2025, starting capital $100,000).

As shown in Fig. 1, the Ensemble model achieved the highest cumulative profit over time, while the Random strategy yielded the lowest, underscoring the effectiveness of machine learning-based predictions in IPO evaluation.

As shown in Fig. 2, IPOs launched during periods of lower inflation (¡2%) achieved the highest long-term success rates, while elevated inflation levels (¿4%) were associated with significantly reduced outcomes.

Fig. 3 illustrates fluctuations in 6-month IPO success rates across years, with a pronounced decline observed in 2021, coinciding with COVID-19 market disruptions.

As shown in Fig. 8, IPOs performed better when the S&P 500 was below its 50-day simple moving average at the time of listing, suggesting contrarian market conditions may favor long-term success.

Fig. 5 reveals seasonal patterns in IPO outcomes, with peak success observed for July launches.

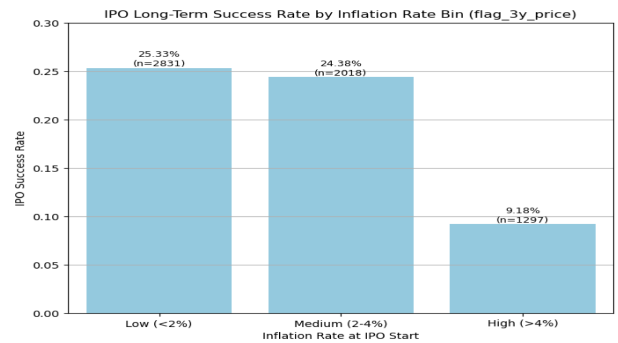

FIGURE 2. IPO 3-year success rates by inflation bin: ˜25% when inflation $<2\%$, ˜24% at 2–4%, and ˜9% when inflation $>4\%$.

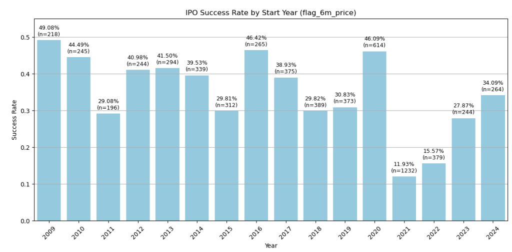

FIGURE 3. 6-month IPO success rate by launch year (2009–2024), with notable dip during COVID-affected 2021.

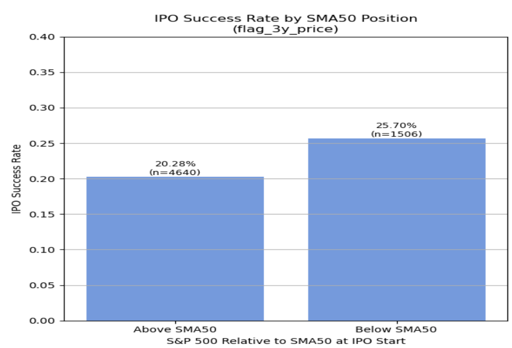

FIGURE 4. Three-year IPO success rates: 20.3% when the S&P 500 is above its 50-day SMA at IPO, versus 25.7% when it's below.

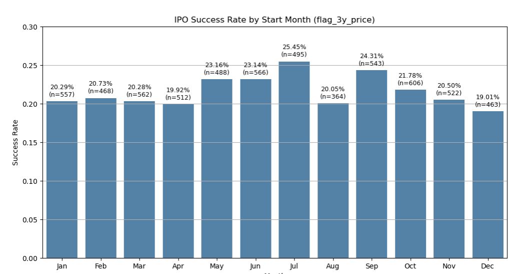

FIGURE 5. Three-year IPO success rate by launch month; highest in July (25.45%) and lowest in December (19.01%).

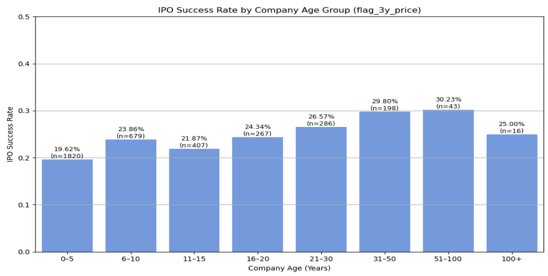

FIGURE 6. Using OpenCorperates data we were able to correlate more establish companies had higher success rates when they opened their IPO

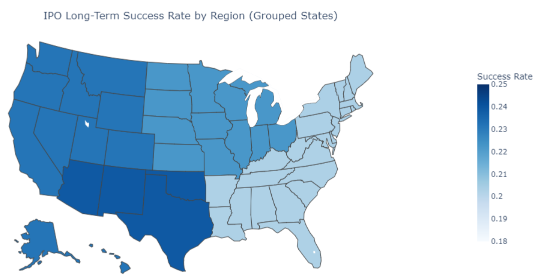

FIGURE 7. We found trends in IPO success rates when looking regionally from east to west cost, with the south west having the highest rate

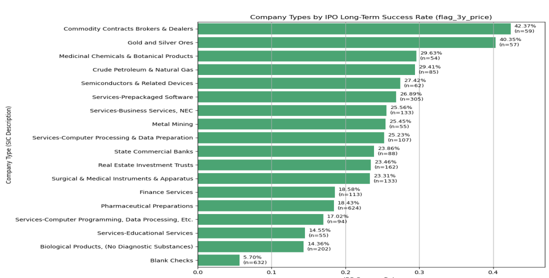

FIGURE 8. Two company types were shown to have the highest IPO success rates being Commodity Contract Brokers and Rare Metal Companies

RECOMMENDATIONS

To enhance the applicability and robustness of the predictive framework, several improvements are recommended. First, future models should incorporate sector-specific segmentation, as IPO success factors differ across industries. Second, integrating unstructured data from S-1 filings and media sentiment using natural language processing (NLP) could improve model accuracy by capturing qualitative risk cues.

Extending the model to international IPO markets would further validate its generalizability. Additionally, deploying the model through a real-time scoring system—such as an interactive dashboard or API—would enable practical application for investors. Finally, regular model retraining is essential to adapt to evolving economic conditions, ensuring continued relevance in periods of macroeconomic volatility.

CONCLUSION

This study demonstrates the viability of machine learning models in forecasting the short- and long-term success of Initial Public Offerings (IPOs). By integrating structured financial disclosures, macroeconomic indicators, and market sentiment variables, the model achieved robust predictive performance. Specifically, the LightGBM model delivered superior results with an AUC of 0.77 for three-year success prediction, outperforming the neural network in both precision and reliability. Feature importance analysis confirmed that key drivers of IPO success include pre-offering financial stability, liquidity metrics, prevailing interest rates, and S&P 500 positioning.

The simulation results further validated the model's economic utility, with the ensemble approach yielding a 40.7% return on investment compared to 6.3% from random selection, yeilding the greatest avg return per trade. These findings confirm the potential of data-driven frameworks to inform IPO investment strategies, reduce risk, and guide capital allocation decisions for financial analysts and institutional investors.

Code Availability

The code and processed data for this study are available on GitHub at: `https://github.com/ghtillem/IPO-Success-Prediction`

ACKNOWLEDGMENT

The authors would like to thank Professor Joshua Loughman for his invaluable guidance, support, and feedback throughout the course of this capstone project. His mentorship greatly contributed to the direction and quality of this research.

REFERENCES

[1] Agrawal, R., and Ananthakumar, U., 2021, "Predicting IPO Underperformance Using Machine Learning," *Decision Sciences Institute Annual Conference*, [Online].

[2] Fischer, T., and Krauss, C., 2018, "Deep Learning With Long Short-Term Memory Networks for Financial Market Predictions," *European Journal of Operational Research*, 270(2), pp. 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

[3] Gu, S., Kelly, B., and Xiu, D., 2020, "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 33(5), pp. 2223–2273. https://doi.org/10.1093/rfs/hhaa009