Decoding Corporate DNA: Unraveling the Predictive Power of Balance Sheet Factors and Macroeconomic Conditions on Long-Term Market Performance

Reet Kothari
University of Washington
Seattle, WA 98195
ryk05@uw.edu

Chase Edson
University of Washington
Seattle, WA 98195
chedson@uw.edu

Braden Reames
University of Washington
Seattle, WA 98195
breames@uw.edu

Tom Tian
University of Washington
Seattle, WA 98195
zhaoyt@uw.edu

Cat Lam
University of Washington
Seattle, WA 98195
tuongcat@uw.edu

Abstract

In the volatile landscape of the stock market, an initial public offering, hereby referred to as IPO, is a critical event for a company that seeks to raise capital and expand its business. However, investing in IPOs is often considered risky and uncertain, as many factors can affect the performance of the newly listed stocks. This paper aims to reduce the skepticism of the common investor by highlighting the most relevant features important to success during the first year after IPO, and showcases the predictive power of these features. We use web scraping to collect data on various company profile metrics and macroeconomic features for 1005 companies that went public between 2017 and 2022. Furthermore, we created visualizations and performed correlation analysis to identify the most influential features for a net gain in an IPO investment. Lastly, we applied different machine learning models, such as Logistic Regression, Support Vector Machine, k-Nearest Neighbor, Adaptive Boosting, Extra-Gradient Boosting, Multi-Layer Perceptron, and Random Forest Classifiers, to predict the one-year net gain of the IPO stocks and evaluate their performance based on accuracy, precision, and return on investment. The paper also provides recommendations for investors based on the findings. The paper concludes by discussing the limitations and future directions of the research.

1 Introduction

Following the recent boom of companies going public, there have been a plethora of IPOs that have come out over the past few years (figure 1). However, market sentiment towards these offerings have been skeptical considering the number of IPOs that tend to fail within their initial year due to bankruptcy, poor performance, or acquisition. To address this issue, and raise the average investor's awareness, we intend to identify and showcase the correlation between a gain on an IPO investment in one year and macroeconomic features such as unemployment rate and average monthly volatility index, as well as company metrics such as total liabilities, stockholders' equity, number of employees, proposed IPO share price, and total volume of shares offered. All of these features were recorded at the time of the company going public for highest relevance. Using these features, we also create predictive models that showcase their application in a simulated investing environment.

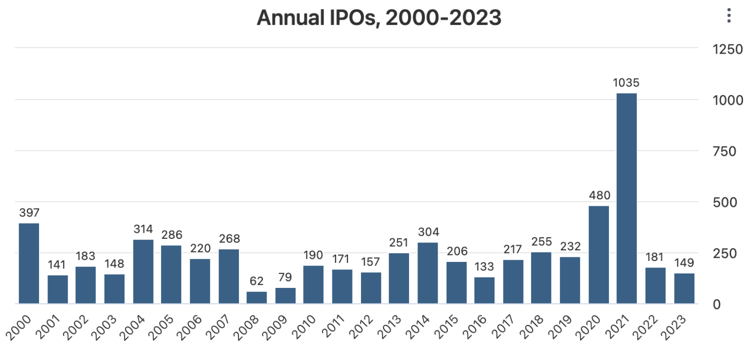


Figure 1: Number of IPOs released every year since 2000

Previous work on this topic tends to rely on stock data to predict future performance. While this approach gives higher accuracy and concrete results, it reduces the utility to the common investor, who may be interested in investing in an IPO at its base price. Therefore, our paper is aimed towards informing the common investor on whether they should invest in an IPO initially, given limited knowledge about the company metrics and the overall economy at the time. This is important for the following reasons:

- **Investment Decision-Making:** The ability to predict the net gain from an IPO investment based on influential features or factors is crucial for investors, financial institutions, and individuals. Despite the vast amount of publicly available data, there has been a gap in leveraging this information effectively for IPO investments. This is primarily due to the complexity of the data and the difficulty in identifying key determinants of success. Our research aims to bridge this gap, making IPO investment decisions more accessible to individual investors who may have limited guidance.
- **Risk Mitigation:** IPO investments are inherently risky, with significant potential for both gains and losses. In 2023 alone, there were 968 IPOs globally, with $101.2 billion in capital raised [7]. Given the scale of these investments, effective risk mitigation strategies are essential. Our research can aid in developing such strategies by identifying the critical features that impact a one-year net gain and providing a confidence interval. This allows investors to better assess and mitigate risks, potentially saving billions of dollars.
- **Company PR:** Companies have an obligation to their shareholders to improve performance and promise great returns on their investment. However, if the current macro or microeconomic situation doesn't favor a bullish attitude, the company could reconsider going public. Our research can provide valuable insights into these situations, helping companies make informed decisions about when to go public.

1.1 Hypothesis

- We hypothesize that balance sheet metrics like Total Liabilities and Stockholders' Equity would highlight the company's performance while they were private and help predict their performance after being public, which would help us to estimate the company's performance after an IPO occurs.
- We hypothesize that a better success prediction can be made with macroeconomic features since that would capture the outside economic environment affecting other companies, which could signify a similar trend for the IPO company.

2 Dataset Description

We web scraped using public APIs as well as using public datasets to obtain the following features for any given company (total set size of 1005 IPOs between 2017 and 2022) and the state of the macro-economy at the time of company IPO, respectively.

Input constructs

- IPO Date: The date that the company went public and launched its IPO
- dealID: The ID for the deal that the company signs to go public, includes company metrics
- Total Assets: A representation of the worth of everything that the company owns at the time of going public
- Total Liabilities: A representation of the combined debts that the company owes at the time of going public
- Stockholders' Equity: A representation of the assets remaining in a company once all liabilities are settled.
- Net Income: A representation of how much revenue exceeds the expenses of a company.
- Number of Shares Offered: The number of shares that the company offers when going public
- Initial Share Price: The base price of an IPO for a single share in the company
- Number of employees: The number of employees working at the company when it goes public
- Unemployment Rate: The unemployment rate during the month of the IPO going public
- Average Monthly Volatility Index (VIX): The volatility index of stocks during the month of the IPO going public
- Price of IPO after 1 year: The future single share price of an IPO, one year after its release.

Output Constructs

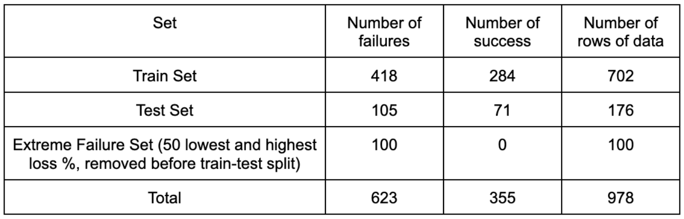
- Success: Binary value determined by if the one year price is greater than the initial base price

The data was scraped using the Yahoo and Nasdaq [10] APIs to get IPO names and dates, and the dealIDs to fill in company metrics. The future price was found using the Yahoo finance library in python [9]. We corrected for reverse splits manually by multiplying the future price by the split ratio. A reverse split is when a company reduces the number of its outstanding shares to increases the price per share proportionally, usually to avoid delisting or to attract more investors. The macroeconomic features like Unemployment Rate and Volatility Index were also found using public datasets (Fred Economic Data [11] and U.S. BLS [12] datasets). The dataset covers key company and macro economic features that could influence the stock performance of a company. The future price and success values help test our model's predictive results.

3 Analytical Approach

We first cleaned up some outliers from the dataset. This included IPOs which were mutual funds (so 0 employees) since those are just subscriptions to shares. We also updated our future price field to handle stock splits and bankrupt companies. This brought down our size from 1340 to 1005 IPOs. Beyond that, we also had to remove some rows for companies, whose net revenue we couldn't find from the deal IDs, bringing the size down to 978. Furthermore, we normalized our features such as total liabilities, Stockholders' Equity, and Net Income by subtracting mean and dividing by standard deviation (so that we can get zero-centered data that our models can distinguish better). For our model we used the following features: Share Price, Shares Offered, Net Income, Stockholders Equity, Total Liabilities, Unemployment Rate, and Volatility Index. The Total Assets field was scrapped since having Stockholders Equity, Liabilities and Assets would only gave 2 degrees of freedom, considering the third value is always dependent on the other 2.

Lastly, we tried different models that scikit offered such as Nearest Neighbor, SVM, ADABoost, Multilayer Perceptron, and Random Forest Classifier. The dataset we procured for the model had a failure to success ratio of 63.7:36.3 (623 rows vs 355 rows). To balance the classess, we removed 50 of the lowest failures (Net loss on investment below 10%) and 50 of the highest failures (Net loss on investment over 85%), making a new set of 100 extreme failures. This led to a split of 59.6 : 40.4 (523 rows vs 355 rows). The 100 rows with the two extremes of a failed IPO investment would be used for further unseen precision testing of the models. The models are judged based on their overall accuracy, precision in finding true IPO failures, and net gain they can attain in our simulated investment environment. The simulated investment environment entails iterating through the test IPO set and determining 2 benchmarks, a naive investor who buys a single stock in every IPO, and a perfect investor who, unrealistically, purchases a single stock in every successful IPO, and makes the highest theoretical gain for single stock investment in IPOs. The naive investor gained **$3.21** and the best investor gained **$575.77**. The model, based on it's success prediction, would also invest one share in the company.



Set	Number of failures	Number of success	Number of rows of data
Train Set	418	284	702
Test Set	105	71	176
Extreme Failure Set (50 lowest and highest loss %, removed before train-test split)	100	0	100
Total	623	355	978

Figure 2: Split of our overall dataset

3.1 Validity

- Our **Construct Validity** is strengthened by several factors. The constructs in this study, including IPO-related metrics and economic features, are well defined - as is the main output construct, success (binary value). The chosen constructs are directly related to the research objectives, and align with the goal of predictiong the success of IPO investments, providing meaningful and relevant insights for investors. The operationalization of these constructs is also well-detailed, as this study explains how each construct was measured or obtained and verified. This transparency ensures the replicability of this study.
- Our **Internal Validity** is maintained through a variety of means as well. We demonstrate significant control over extraneous variables, including the normalization of features and the removal of IPOs that were not beneficial to the study. The detailed account of the data cleaning procedures, such as correcting for stock splits and bankrupt companies demonstrates the internal integrity of the data, as does the removal of incomplete or otherwise invalid rows. Finally, the recognition of class imbalance in the failure to success ratio, and the steps taken to balance the classes indicate strengthened internal validity.
- Our **External Validity** is improved through the inclusion of a diverse dataset covering IPOs between 2017 and 2022. This timeframe and sample size enhance the generizability of the findings to a broader population of IPOs. Additionally, the use of public datasets contributes to external validity. Finally, the approach of simulating the trading strategies of predictive models enhances external validity.

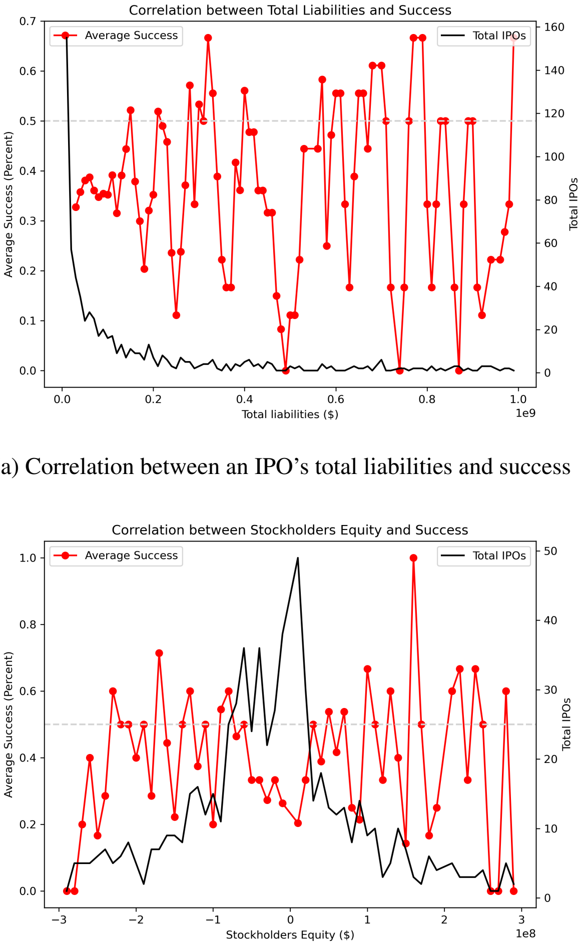
3.2 Evaluation and Success

We would evaluate our approach based on key performance indicators that we have set for our models:

- For the prediction scores, we are aiming for a model over 60% and preferable over 70%. This number was determined by the typical success ceiling for expert investors [8].
- We also want to put an emphasis on ensuring that the model predicts failed IPOs better than successful ones in an attempt to improve risk management. We are aiming for an accuracy of 70-80% on the extreme failure set.
- As for the model's investment decisions, we want the gain to be about double the gain of a naive investor and more than half the gain of the best investment strategy.

4 Results

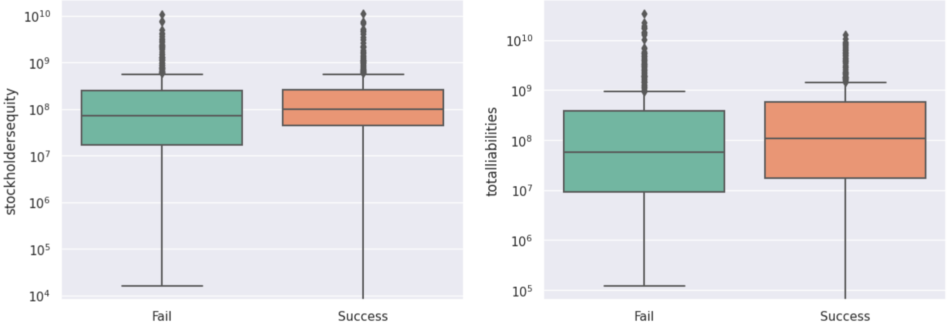
In our initial data analysis, it became clear that many of the balance sheet metrics tested did not have an obviously measurable impact on the performance of a given company's value one year after the IPO took place. In figure 3, no relationship between the overall rate of success, and total liabilities and stockholders' equity is visually apparent. This could be because of the high amount of IPOs concentrated around 0.0 - 0.1e9 total liabilities and stockholders' equity values.



(a) Correlation between an IPO's total liabilities and success

(b) Correlation between an IPO's stockholders' equity value and success

Figure 3: Testing first hypothesis

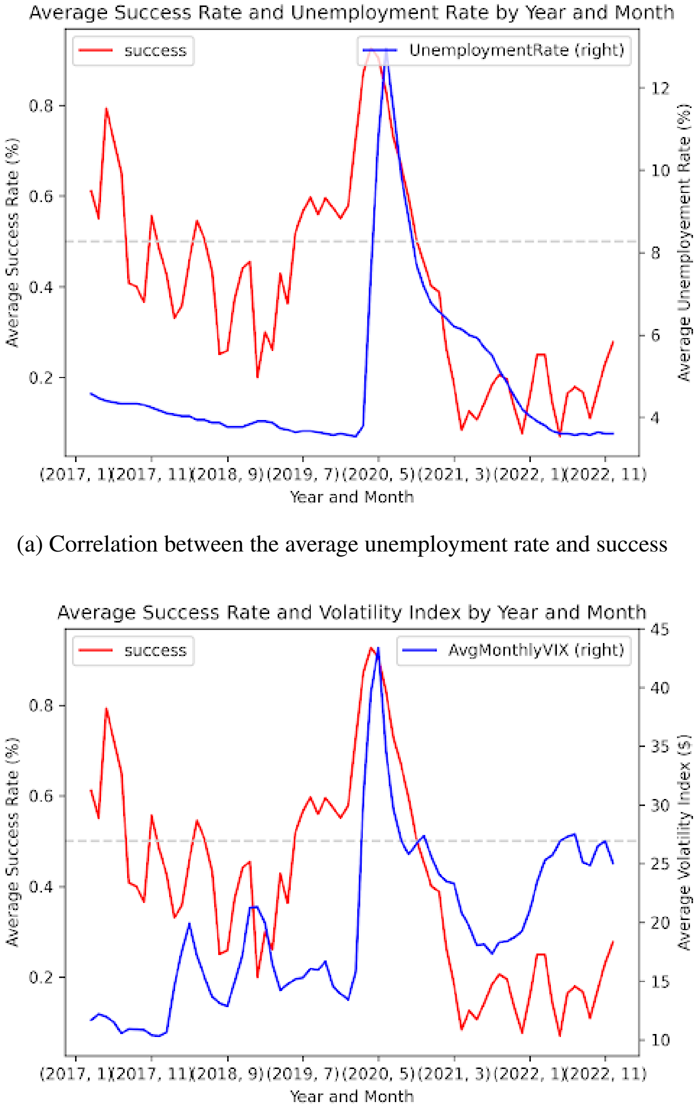


(a) Boxplot for distribution of stockholder equity based on success

(b) Boxplot for distribution of total liabilities based on success

Figure 4: Average balance sheet value for successful and unsuccessful IPOs

However, as can be seen in figure 5, unemployment rates and market volatility have a somewhat direct correlation with the success of an IPO (notably during to the COVID-19 pandemic in 2020), though it seems to show a somewhat lagged effect where the volatility and unemployment rates falls or rises after the actual effect takes place on the average success rate.

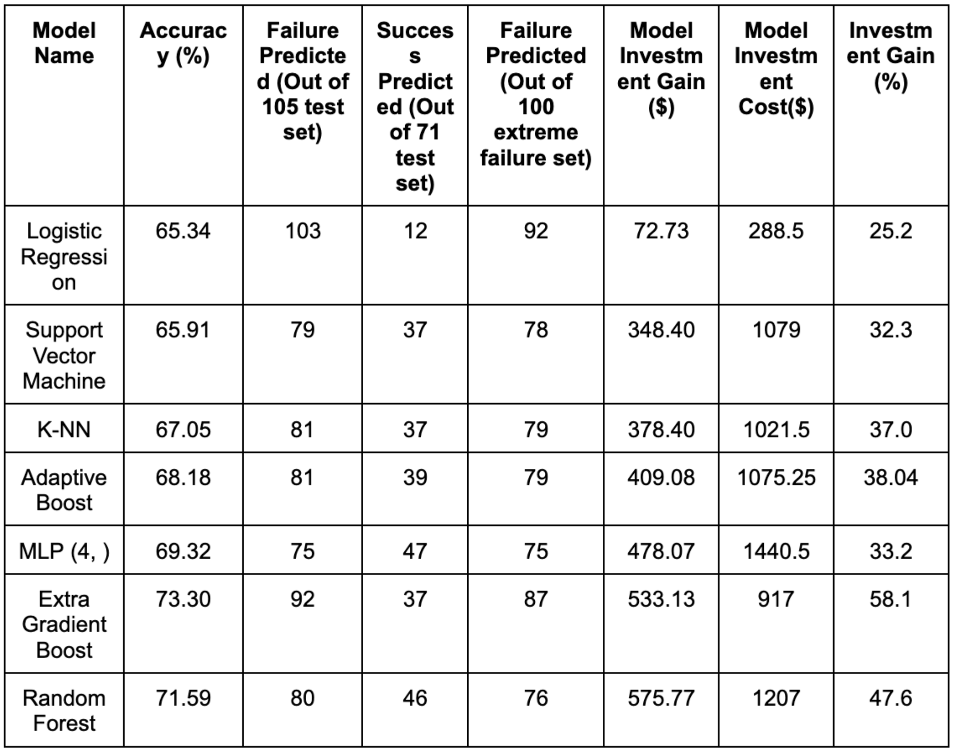


(a) Correlation between the average unemployment rate and success

(b) Correlation between the average market volatility and success

Figure 5: Testing second hypothesis

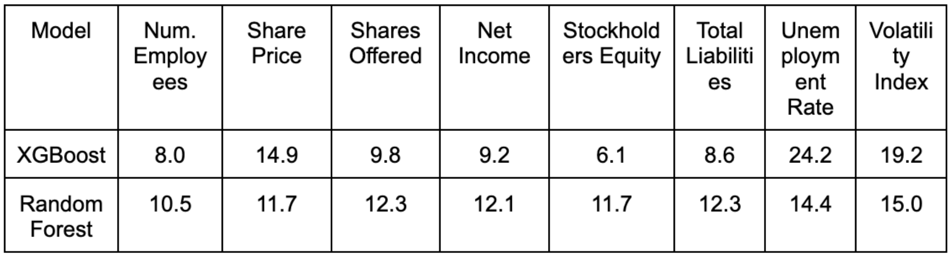
Due to lack of obviously linear relationships, we employ a few linear and non-linear models on the same task to recognize the predictive power of such features. As highlighted in figure 6, there is a clear shift between our Logistic Regression Classifier and our Multi-layer Perceptron or Random Forest Classifiers. The linear models seem to predict failures a lot better than the non-linear ones, but this could be due to overfitting on to the labels, as reflected by their poor precision on successful IPOs (and gain in investment environment).



Model Name	Accuracy (%)	Failure Predicted (Out of 105 test set)	Success Predicted (Out of 71 test set)	Failure Predicted (Out of 100 extreme failure set)	Model Investment Gain ($)	Model Investment Cost($)	Investment Gain (%)
Logistic Regression	65.34	103	12	92	72.73	288.5	25.2
Support Vector Machine	65.91	79	37	78	348.40	1079	32.3
K-NN	67.05	81	37	79	378.40	1021.5	37.0
Adaptive Boost	68.18	81	39	79	409.08	1075.25	38.04
MLP (4,)	69.32	75	47	75	478.07	1440.5	33.2
Extra Gradient Boost	73.30	92	37	87	533.13	917	58.1
Random Forest	71.59	80	46	76	575.77	1207	47.6

Figure 6: Predictive Power and Investment Results of our models

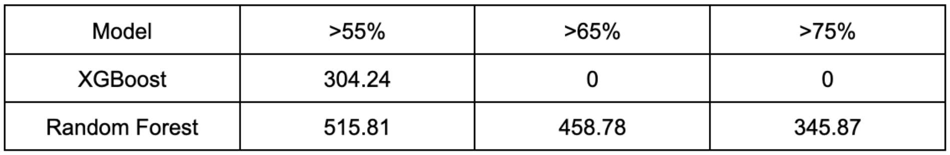
For the two best models, with XGBoost being a supervised model, we can see the feature importance that they give for the features in our dataset in figure 7. We can see clearly that both models tend to show a preference towards macroeconomic factors to make their predictions. Moreover, there seems to be a high importance given to the combination of share price and shares offered, which may highlight that accessibility of the IPO to a bigger audience (low price and high volume), might lead to a better chance of IPO success.



Model	Num. Employees	Share Price	Shares Offered	Net Income	Stockholders Equity	Total Liabilities	Unemployment Rate	Volatility Index
XGBoost	8.0	14.9	9.8	9.2	6.1	8.6	24.2	19.2
Random Forest	10.5	11.7	12.3	12.1	11.7	12.3	14.4	15.0

Figure 7: Feature Importance (in %) values for the best models

To evaluate the effectiveness of the XGBoost and Random Forest classifiers in guiding stock investment decisions, we analyzed their prediction confidence levels. We defined specific confidence ranges and observed their impact on investment choices. The analysis revealed that while XGBoost boasts higher overall accuracy, it tends to be conservative in making investment decisions, often showing modest confidence in its predictions. This trait is evident in Figure 8, where XGBoost shows a reluctance to make investments in the higher confidence ranges. Such cautious behavior suggests that XGBoost may be more suitable for risk-averse investors, as it demonstrates a strong ability to predict failures while maintaining respectable performance even with modest confidence levels. In contrast, the Random Forest classifier exhibits tighter confidence bounds, indicating a higher certainty in its predictions. This model displays a wider variation in confidence across its investment decisions, potentially appealing to risk-tolerant investors who are open to a broader spectrum of opportunities, including those with the potential for larger gains.



Model	>55%	>65%	>75%
XGBoost	304.24	0	0
Random Forest	515.81	458.78	345.87

Figure 8: Investment gains when success confidence is greater than particular range(gains are cumulative over the confidence range)

4.1 Discussion of Results

From our data visualizations, and model feature importance, we can see that the macroeconomic conditions seem to give the best indication for the success of an IPO investment in a one-year time span. This suggests that IPOs move along with the general market and sway in a similar direction as the volatility index of the market.

As for our models, considering the disparity between the logistic regression classifier and random forest classifier accuracies, non-linear models would have a better chance of predicting the success of an IPO. Of the non-linear models, the XGBoost model predicts the highest amount of failures while also ensuring high investment gains. This indicates that the model would be suited for a safe investor who is more focused on avoiding failed IPOs. This is further supported by the confidence interval investment choices that the model showcases.

On the other hand, an investor who's interest in making the highest gain, and has some more risk tolerance, could probably consider the random forest classifier since that has one of the highest precisions for successful IPOs while also ensuring a high failure precision.

The feature importance for the XGBoost model seems to showcase a relatively low importance for stockholders' equity and total liabilities. This could indicate that they aren't as relevant in determining the failure of an IPO. Surprisingly, the share price and shares offered have a high importance, implying that there might be an unexplored correlation there.

5 Related Work

We review some of the previous works that have applied machine learning techniques to predict the performance of IPOs. We also highlight the main differences and contributions of our approach compared to these works.

- **Predicting IPO initial returns using random forest**: This paper proposes a method to predict the initial returns of IPOs using random forests, a non-linear ensemble learning technique. The authors benchmarked this algorithm against eight classic machine learning algorithms and found that random forests outperformed the alternatives in terms of mean and median predictive accuracy [5]. We took inspiration from this work to try a random forest classifier and found that, with different features, the performance of the model can be improved further.

- **Predicting IPO Underpricing with Genetic Algorithms**: This paper introduces a rule system to predict the first-day returns of initial public offerings (IPOs) based on the structure of the offerings. The solution is based on a genetic algorithm using a Michigan approach. The performance of the system is assessed by comparing it to a set of widely used machine learning algorithms. The results suggest that this approach offers significant advantages on two fronts: predictive performance and robustness to outlier patterns. The importance of the latter should be emphasized as the results in this domain are very sensitive to their presence [6]. Due to the short-sightedness of their success time frame, and different feature set, we believed this paper to be useful in understanding the problem better but not necessarily in line with our approach.

6 Conclusion

The findings from our study demonstrate the complexity of evaluating initial public offerings as part of an investment strategy, and provide valuable insights for common investors and financial institutions alike, as well as companies considering going public. Our research aimed to address the skepticism surrounding IPO investments, and specifically focused on identifying the key determinants of success within the first year of a company going public. Our initial analysis demonstrated that traditional balance sheet metrics, such as total liabilities and stockholders' equity, did not exhibit an obvious linear relationship with the success of an IPO one year after going public, while macroeconomic factors like unemployment rates and market volatility demonstrated a more direct correlation with IPO success.

To determine the power these features held in making educated predictions, we employed various linear and non-linear models. Notably, our research showcased a significant difference in performance between linear models like Logistic Regression and non-linear models, such as Random Forest Classifiers and Multi-layer Perceptron. Feature importance analysis further indicated a preference in our models towards macroeconomic factors, emphasizing their role in predicting IPO success. The evaluation of XGBoost and Random Forest classifiers additionally showed their suitability for different investor profiles. XGBoost demonstrated a more conservative approach, suitable for more risk-adverse investors. The Random Forest classifier contrasted this approach with tighter confidence bounds, appealing to more risk-tolerant investors who are open to a broader range of outcomes.

While our research provides valuable insights for investors, it is still essential to acknowledge its limitations. Future research could further explore additional factors influencing IPO success in addition to further refining these predicitve models. Additionally, considering the dynamic nature of financial markets and the overall economy, ongoing data enhancement could improve this approach over time. Our findings contribute to the evolving landscape of IPO investment strategies, leading to further development in this field.

7 Ethical Considerations

In this project, we aim to develop a model that can predict the success of IPOs based on various features, such as company metrics, macroeconomic indicators, and market sentiment. Our goal is to provide insights for investors, financial institutions, and companies who are interested in IPOs investments. However, we also acknowledge that there are some ethical issues and challenges associated with our project, such as:

- **Data Quality and Reliability**: We rely on web scrapping and public APIs to collect our data, which may introduce errors, biases, or inconsistencies in the data. We perform data cleaning, outlier detection, and normalization techniques, and we manually adjusted for stock splits and reverse splits. We also disclose the sources and limitations of our data in this paper.
- **Model Fairness and Accountability**: Our model may not capture all the factors that influence the success of IPOs, and it may produce inaccurate or misleading predictions for some cases. We evaluate our model using various performance metrics, such as accuracy, precision, recall, and net gain. We also report the feature importance and confidence scores of our model, and we provide disclaimers and caveats for our predictions.
- **Social and Environmental Impact**: Our project may have social and environmental implications, such as affecting the allocation of capital, the distribution of wealth, the innovation and growth of companies and the sustainability and well-being of society. We suggest that our model's predictions should not be the sole basis for IPO investments, but rather one of the many factors to consider.

Disclaimer: It is important to stress that our model and its predictions are **not intended to serve as a definitive guide for investors**. Instead, they should be viewed as **a reference point or an additional tool** that can supplement an investor's own research and analysis. Investment decisions are complex and should take into account a wide range of factors, many of which may not be included in our model. Therefore, while our project can provide valuable insights, it should not replace professional financial advice or personal judgment. We strongly recommend that investors consider our findings as one of many resources, and that they consult with a financial advisor or conduct their own thorough research before making investment decisions. In conclusion, our project aims to contribute to the broader conversation about IPO investments, but it does not offer a foolproof road map to success. It is, after all, only a reference.

7.1 Transparency and Explainability

We need to ensure that there is transparency in how our model was developed by making it (and the process for creating it) understandable to stakeholders. We need to ensure that the output of the model should not be considered as fact, especially in financial contexts where trading decisions can have significant consequences. For this, we need to clearly communicate the limitations of this model, and the potential for errors that exists.

7.2 Bias and Fairness

We need to be vigilant about presenting all biases that may exist within the model, and the data used to create it, as these biases can lead to unfair or unpredictable predictions, especially in unseen scenarios. We also need to consider the full impact of this model, especially across different demographics, and strive for fairness on its' impact.

References

[1] IPO statistics and charts. (2023, December 4). https://stockanalysis.com/ipos/statistics

[2] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[3] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[4] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[5] Baba, B. & Sevil, G. (2020, March) *Predicting IPO initial returns using random forest.* Retrieved December 10, 2023, from https://www.sciencedirect.com/science/article/pii/S2214845019302686#bib57.

[6] Luque, C., Quintana, D. & Isasi, P. (2012, March) *Predicting IPO Underpricing with Genetic Algorithms*. Retrived December 10, 2023, from https://davidquintana.apps-1and1.net/wp-content/uploads/2017/11/Predicting-IPO-Underpricing-with-Genetic-Algorithms.pdf.

[7] Gerringm, R., Choi, R. & Steinbach, M. (2023, September 27) *As the market starts to shift, how can your IPO be built to last?.* Retrieved December 10, 2023, from https://www.ey.com/en_gl/ipo/trends.

[8] Sajumon, A. (2023, September 14). How to predict stock price for next day?. Fisdom. https://www.fisdom.com/predict-stock-price-for-the-next-day/

[9] Ranaroussi. (2023, December 6). Ranaroussi/yfinance: Download market data from Yahoo! Finance's API. GitHub. https://github.com/ranaroussi/yfinance

[10] Time-series. Nasdaq Data Link Documentation. (n.d.). https://docs.data.nasdaq.com/docs/time-series

[11] Federal Reserve Economic Data: Your trusted data source since 1991. St. Louis Fed Web Services: FRED® API. (n.d.). https://fred.stlouisfed.org/docs/api/fred/

[12] U.S. Bureau of Labor Statistics. (n.d.). Databases, tables amp; calculators by subject. U.S. Bureau of Labor Statistics. https://www.bls.gov/data/unemployment