Accurately Predicting IPO Underpricing Using Machine Learning

Lisa Warren Marcell Borhi Abe Leininger Evan Jackson

1 Problem Space

The main objective of the project is to predict the short-term returns of IPOs in the United States through the creation of a random forest model that determines whether a given IPO will be underpriced. IPO underpricing describes the increase in stock price from the initial offer price to the first-day closing price. If the IPO offer price is too high relative to the fundamental value, investors will not invest. If the IPO price is too low compared to its fundamental value, then the issuing firm leaves money on the table. The project will take into consideration many variables, from key financial ratios, investor conditions at the time of the IPO, and key stakeholders in each transaction for historical IPOs in the last few decades. This project was expected to be constrained by the process of retrieving quantitative (such as key financial ratios) and qualitative data (such as market sentiments at the time of the IPO). It is also constrainted by the information asymmetry created by the absence of publicly available data for pre-IPO firms. One challenge we expected to face in this project is the uncertainty surrounding the influence of the principal-agent problem and other unquantifiable variables surrounding IPOs.

We were unable to find many historical attempts at predicting short-term IPO returns in the United States, however we were able to source studies by Agrawal et al. and D. Meng that attempted to predict short-term IPO returns in other markets such as India and China respectively (2021, 2008). From our research, historical attempts have focused on different factors, such as key financial ratios or the content of the IPO prospectus' in historical IPOs. Different variations of machine learning techniques used in existing research include generalized linear models, gradient boosting models, neural networks, and natural language processing. For our project, we drew from existing research studies and preliminary research on the financial markets to determine the factors that we decide to implement in our model.

2 Data Collection and Processing

Before beginning data collection, we created a list of features we wanted to include in our model. These features included offer price, market capitalization, the percent of shares being sold, industry sector, sales growth, offer size, GDP, unemployment rate, and a few other characteristics about the firm, offering, and economic conditions. These were determined based on what features we thought would be useful and what features were used by previous researchers. Our goal was to find data for as many of the features as possible, but we knew that we would be limited by what data was available.

Our initial data came from IPOScoop and Ticker.com. We collected the data from IPOScoop from a publicly available CSV file that included all of the IPOs (around 3500) that they had covered along with their opening price, closing price, and lead manager. The Ticker data was gathered using pandas to pull the data straight from the website. This data set included the features from IPOScoop along with the sector, offer size, turnover, and gain. While we used these sources to create an initial random forest model, they lacked many features that we wanted to consider. Ultimately, we ended up collecting all of the offering and firm data from the Bloomberg database. The data was collected using an advanced search for all IPOs from 2000 to 2022 that was exported to a CSV file. Bloomberg allows you to select a variety of columns, so we were able to get just about every feature relating to the offer and the company that we wanted. We gathered quarterly economic data using the FRED API offered by the St. Louis Fed. The St. Louis Fed archives a variety of economic indicators, such as real GDP per capita, OECD leading indicator, interest rate, seasonally adjusted unemployment rate, and CPI growth rate. The table with all the features we considered is listed below.

Feature	Definition
Sales - 1 Yr Growth	The percent growth of the company's sales in the year before its initial public offering
Profit Margin	The company's profit margin (percent of each dollar of sales that is profit) prior to initial public offering
Return on Assets	The company's return on assets (the company's net income divided by total assets) prior to initial public offering
Offer Size (M)	The Offer Price multiplied by the number of Common Shares sold in the Share Offering
Shares Outstanding (M)	The number of shares released to the public (in millions)
Offer Price	The price of a single share of the IPO at the start of the first day it is publicly traded
Market Cap at Offer (M)	Shares outstanding multiplied by the offer price
Cash Flow per Share	The company's "after-tax earnings plus depreciation" per share
Instit Owner (% Shares Out)	The percentage of total shares that are not owned by the company
Instit Owner (Shares Held)	The number of shares held by the company
Real GDP Per Capita	The value of the final goods and services produced in the United States
OECD Composite Leading Indicator	Designed to provide early signals of turning points in business cycles showing fluctuation of the economic activity around its long-term potential level. Shows short-term economic movements in qualitative rather than quantitative terms
Interest Rate	The interest rate at which the central bank lends to commercial banks to meet their liquidity needs
Seasonally Adjusted Unemployment Rate	The number of unemployed as a percentage of the labor force, adjusted for normal seasonal changes
CPI Growth Rate	Growth rate from previous period for the Consumer Price Index including All Items
Industry Sector	The broad industry sector the IPO falls into represented as a number from 0 to 9.
Industry Group	The group within its sector that the IPO falls into represented as a number from 0 to 63.
Industry Subgroup	The subgroup within its group that the IPO falls into represented as a number from 0 to 264.
Offer To 1st Close	The percent change from the offer price to the stock's price at the end of the first day of trading. Used to define the underpriced parameter
Underpriced	Label parameter. 1 if the IPO is underpriced and 0 if the IPO is not

The data required processing before it could be plugged into our models. We began by combining the macroeconomic and IPO-specific data. This was done by determining the quarter in which each IPO went public and concatenating the economic indicators for that quarter using pandas. From here, the data was cleaned by removing any IPOs that were missing data, as Bloomberg did not have certain firm data was not given for some IPOs, and converting textual features, such as industry sector, to numbers. We ended up with approximately 1700 IPOs, about 72% of which were underpriced, and 17 features.

3 Techniques/Algorithms

We implemented a random forest model to determine whether an IPO will be underpriced. The random forest algorithm works by creating multiple decision trees from a subset of the data. Each tree is built using a different set of features, and each tree is used to make a prediction. The final prediction is determined by taking the majority vote of all the individual tree predictions. This process helps reduce overfitting and improves accuracy because it reduces the variance in the model. Random forest also has the ability to handle large datasets with high dimensionality, making it a powerful tool for predictive analytics. By randomly selecting subsets of features, random forest can identify important features that may have been overlooked in

traditional methods. Additionally, random forest can identify non-linear relationships between variables that may not be apparent in other models. Finally, because it randomly samples the dataset, random forest can handle missing values and outliers better than other algorithms, making it more robust and reliable. This was an important characteristic for our model as consistent and reliable data can occasionally be difficult to source for IPOs, especially when it is qualitative.

The time complexity of random forest is $O(n \cdot m \cdot log(m))$, where n is the number of data points and m is the number of features. The space complexity is $O(n \cdot m)$, where n is the number of data points and m is the number of features.

The main limitation of random forest is that it can be computationally expensive. Additionally, it can be difficult to interpret the results of a random forest model, as the individual decision trees are not easily interpretable. Finally, random forest can be prone to overfitting if the number of trees is too large or if the data is noisy.

An alternative option to random forest would have been to implement a neural network. In contrast to random forests, neural networks are better suited for datasets with fewer features and lower dimensionality. They are also better at identifying complex non-linear relationships between variables that may not be apparent in other models. Given that our data has many features and there is not necessarily a non-linear relationship between the factors which we used, using a random forest model is the best option. This claim is also supported by numerous research papers (Agrawal & Ananthakumar, 2021). Another alternative to the random forest algorithm would have been a gradient boosting model. Gradient boosting is a machine learning technique that involves combining multiple weak models to create a strong predictive model. This is typically done by adding models to the ensemble one at a time, with each model being trained to correct the errors made by the previous models. The final model is a weighted average of all the individual models, with the weights determined by how well each model performed. This technique is often used with decision tree models, and can be used for both regression and classification problems. Lastly, we considered an alternative model, the support vector machine. This algorithm separates data points into classes using a hyperplane, and can be used for classification and regression problems.

4 How Solution Models Human Thought Process

For financial analysts, evaluation of attractive prospects for IPOs typically focuses on a few main characteristics. These include, but are not limited to, the unique value proposition of the company's business model, a differentiated and unique product, a strong revenue growth and margin expansion, and a competent management team. Teams within investment banks may use features like these to decide whether or not to further investigate the prospect of an IPO for the chosen company. If they decide to investigate the prospect further, they will conduct a deeper analysis of other company characteristics, such as their industry competitors, internal capabilities, financial metrics, and other information that may not be available to the public. This process largely follows the structure of a decision tree - some factors are more determinant of whether or not a company pursues an IPO than others. For example, the presence of a strong management team or the size of the total addressable market may be a factor at the 'top' of the decision tree. Similarly, our random forest model implements a number of decision trees to reduce overfitting and improve the accuracy of the prediction.

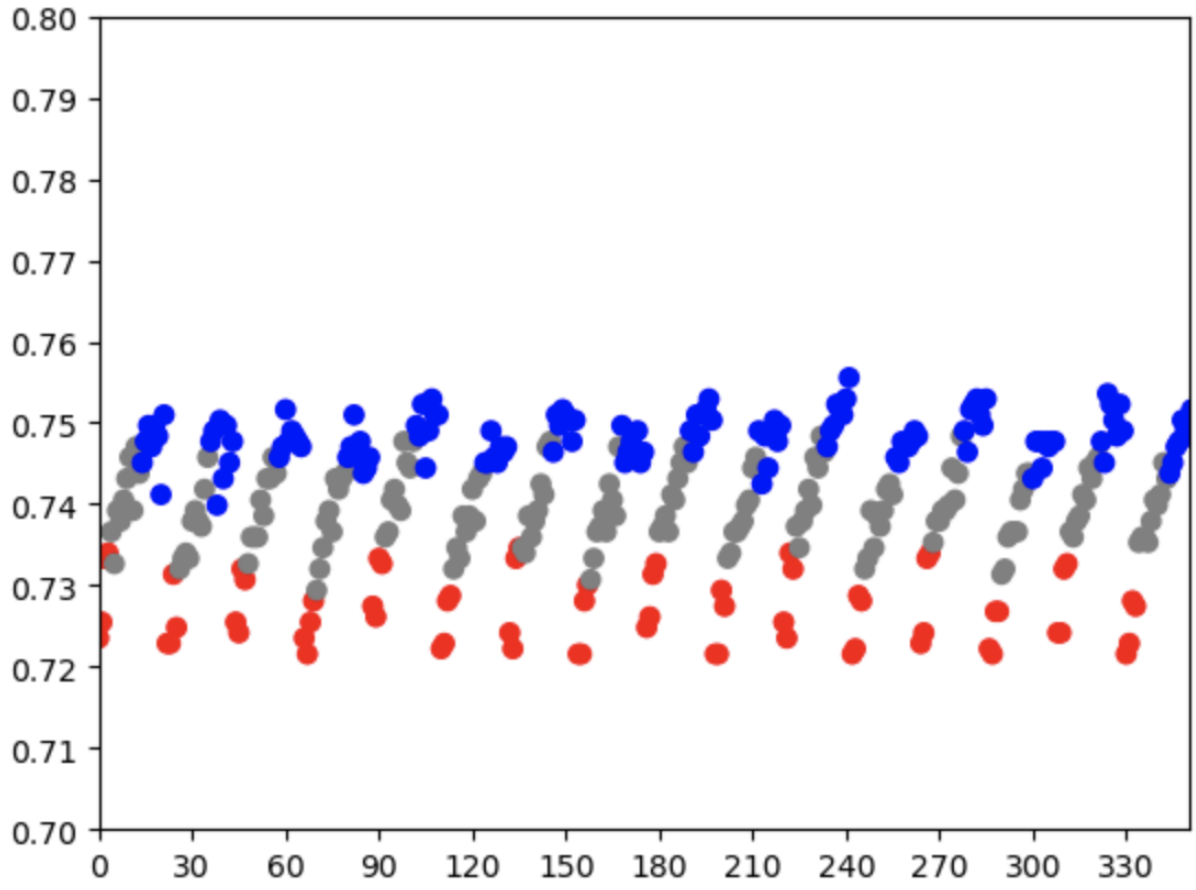
5 Results

Based on what data was available and what data was used by other researchers, we decided to run our random forest model using the following features: the company's sales, profit margin, return on investment, and cash flow in the year before going public, offer size (the offer price multiplied by the number of shares made available to the public), the number of shares outstanding (the number of shares not held by the company itself), the offer price, the market capitalization (the market price of all outstanding shares), the percent of shares

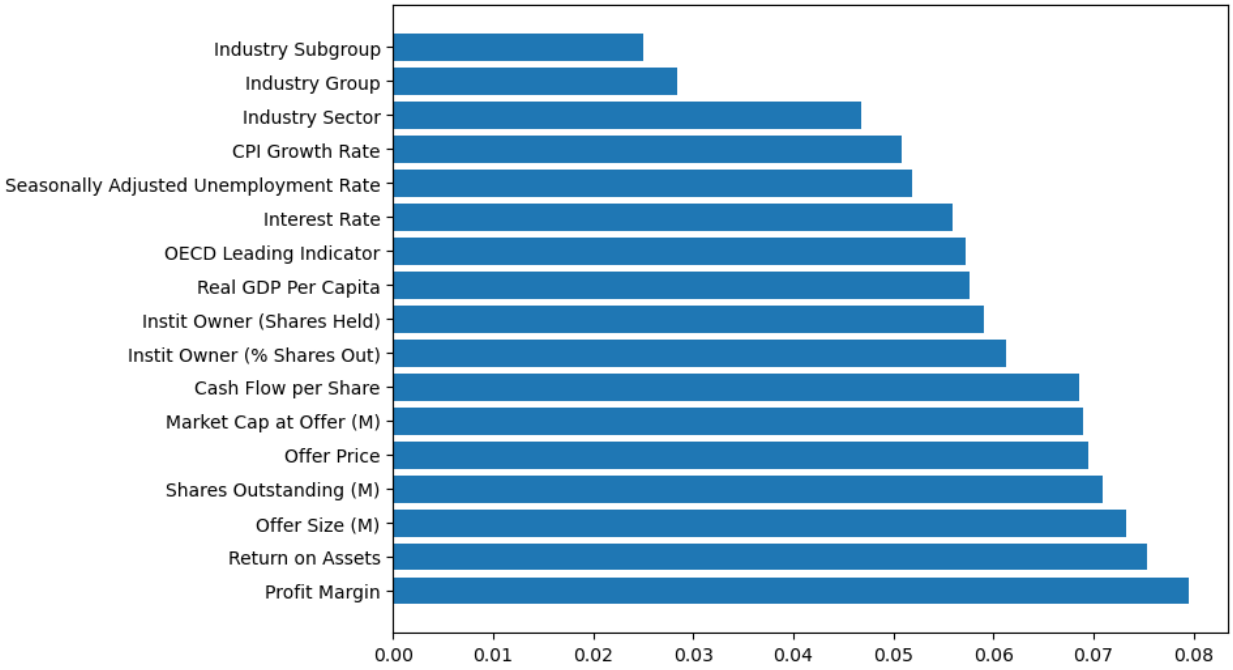
that were outstanding, the number of shares that were held by the company, industry sector, and industry group. We also included macroeconomic indicators such as the OECD leading indicator for the U.S., the U.S. interest rate, seasonally adjusted unemployment rate, and the Consumer Price Index (CPI) growth rate.

To find the most optimal configuration for the random forest model, we trained a permutation of several parameters: the number of estimators, the maximum depth of the tree, the criterion for splits, and the maximum number of features to consider. After training all of the models over the same data set, three times each and averaging results, we found that the models with a large max depth perform significantly better. Moreover, the model using 600 estimators, considering log2 of the total number of features, determining splits using Gini impurity, and having a max depth of 14, had the best results.

The graph below details the positive correlation between accuracy of the model and max depth of the model. Blue points on the graph represent models with a max depth over 10, grey points represent models with a max depth between 5 and 10, and red points represent models with a max depth below 5.



The model is able to accurately predict whether an IPO will be underpriced 75.53236% of the time.



From the results, we can see that the most pertinent features appear to be the company's characteristic features — profit margin and return on assets — followed by the information about the actual offering such as offer size, shares outstanding, and offer price. This is interesting because these are features that would naturally be considered by prospective investors, showing further how our model replicates human thought processes.

These results were actually a bit better than expected. 75% is a relatively high accuracy rate for any machine learning model, and it is especially high for IPOs, which are generally considered to be unpredictable. Previous literature has reported similar accuracy rates using random forest models to predict IPO underpricing, but given our relatively small dataset, we did not necessarily expect to come as close as we did. The accuracy of the model implies that there are additional factors to consider when predicting the pricing of an IPO. This could include textual data, economic and market factors, sentiments surrounding the company, or other company performance data that was unavailable to us. However, it is very likely that, even with better data, our accuracy could not get higher than 85-90% because the stock market is so hard to predict.

Breaking down the random forest results by class, our model achieved a class accuracy of 92.4% for underpriced IPOs and 22.6% for non-underpriced IPOs. In comparison to our secondary models (gradient boosting, SVM, and neural network models), the random forest model performed as follows:

Model	**Description**	**Accuracy**
Random Forest	Combination of multiple decision trees.	76%
Gradient Boosting Classifier	Ensemble of weak decision tree models.	75.3%
Support Vector Machine	Finds hyperplane that best separates different classes.	73.9%
Binary Classification Neural Network	Adjust weights to minimize the loss	70.2%

Ultimately, our results imply that predicting whether an IPO will be underpriced may be very feasible with publicly available data. However, the results also imply that a more useful tool for the financial services sector might be a model that predicts the extent to which an IPO is under or overpriced. Additionally, our model and our understanding of the capital raising process for companies wishing to go public imply that the accuracy of such models may be increased with the implementation of private, IPO-related data.

How This Project Could be Improved

As mentioned previously, the biggest bottleneck in our project is the availability of data. It is very difficult to find information on several firm characteristics. Specifically, there is no central database containing firms' profit margin for the year prior to the IPO listing, asset turnover ratio and return on assets prior to the listing, the debt to asset ratio, and total assets. This is due to the fact that generally, companies who IPO typically do not have a long history of disclosing their financial information, something that is only legally required once a company goes public. Including these features in our random forest model may improve its accuracy. Additionally, factors such as the firm's management team and track record also influence the success of an IPO listing, and thus should also be taken into account for the most accurate prediction. However, it is very hard to assign a firm's management team a comprehensive numerical value. Similarly, it is extremely difficult to compare various firms' track records to come up with an accurate scale.

Initial valuation and pricing of the IPO is also heavily dependent on the methods used by the investment bank advising the company at the center of the offering. Throughout valuation, investment banks create models based on assumptions they have made about future growth estimates of the company. However, it is difficult to access the actual valuation process used by investment banks to price IPOs - this is not something that the company is required to disclose when going public. We believe that, given data surrounding the valuation process and the assumptions on which valuation operates, our random forest model may be able to more accurately classify IPO short-term performance. However, after experiencing difficulties in sourcing accurate, relevant data for this project, we acknowledge that this may be impossible to complete while operating externally to the investment banks who work on these transactions. As long as details regarding the valuation process are kept private, we believe that this may be a research area of interest for internal teams of investment banks who are looking to use data to more effectively price IPOs.

Another way we could potentially improve our model would be to implement Natural Language Processing (NLP) to gather additional data from the companies' S-1 Filings, like Katsafados, et. al do in their paper. The S-1 filing is a form that businesses are required to fill out before going public, containing business information such as "business operations, the use of proceeds, total proceeds, the price per share, a description of management, financial condition, the percentage of the business being sold by individual holders and information on the underwriters" to help inform investors about the IPO (Investopedia). Since investors frequently use these forms when deciding whether or not to invest in an IPO, being able to process the data held within could help the machine learning model even more closely replicate human thought processes and more accurately predict what IPOs will be under or overpriced. We initially considered utilizing NLP for our model, but we determined that, given the complexity of NLP, it wouldn't be feasible for us to implement in just a few short weeks. If this were expanded into a larger project, however, adding this data could be an effective next step to making our model more accurate and making up for the lack of firm characteristic data that is currently the project's primary pitfall.

Works Cited

Agrawal, Rachit & Ananthakumar, Usha. (2021). Predicting IPO underperformance using machine learning.

D. Meng, "A Neural Network Model to Predict Initial Return of Chinese SMEs Stock Market Initial Public Offerings," 2008 IEEE International Conference on Networking, Sensing and Control, 2008, pp. 394-398, doi: 10.1109/ICNSC.2008.4525247.

Katsafados, A. G., Androutsopoulos, I., Chalkidis, I., Fergadiotis, M., Leledakis, G. N., & Pyrgiotakis, E. G. (2020, October 27). Textual Information and IPO Underpricing: A Machine Learning Approach.

Kenton, Scott. "SEC Form S-1: What It Is, How to File It or Amend It." Investopedia

Wang, K. (2021). A comprehensive study of the reasons for underpriced IPOs. Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021).

International Monetary Fund, Interest Rates, Discount Rate for United States [INTDSRUSM193N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/INTDSRUSM193N, December 6, 2022.

Organization for Economic Co-operation and Development, Consumer Price Index: Total All Items for the United States [CPALTT01USM657N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPALTT01USM657N, December 6, 2022.

Organization for Economic Co-operation and Development, Leading Indicators OECD: Leading indicators: CLI: Normalised for the United States [USALOLITONOSTSAM], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/USALOLITONOSTSAM, December 6, 2022.

U.S. Bureau of Economic Analysis, Real gross domestic product per capita [A939RX0Q048SBEA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/A939RX0Q048SBEA, December 6, 2022.

U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, December 6, 2022.

Warrington College of Business. (April 28, 2022). Share of IPOs with negative first day earnings per share (EPS) in the United States from 1980 to 2021 [Graph]. In Statista. Retrieved December 07, 2022, from https://www.statista.com/statistics/429868/share-of-ipo-deals-with-negative-first-day-return-usa/