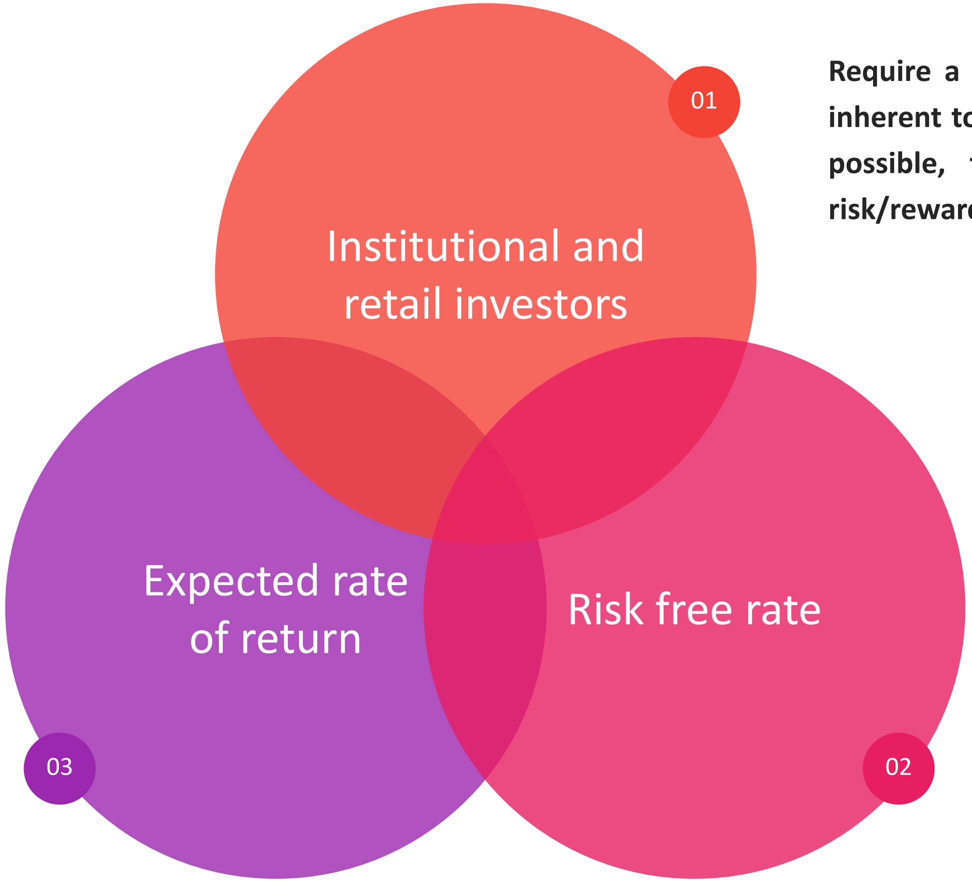
HyIPO

HyIPO: Hyped Initial Public Offerings

Business Problem

Build a model which ranks IPOs in terms of their expected returns

Require a tool for asserting the risk inherent to these investments and, if possible, to select only the best risk/reward combinations



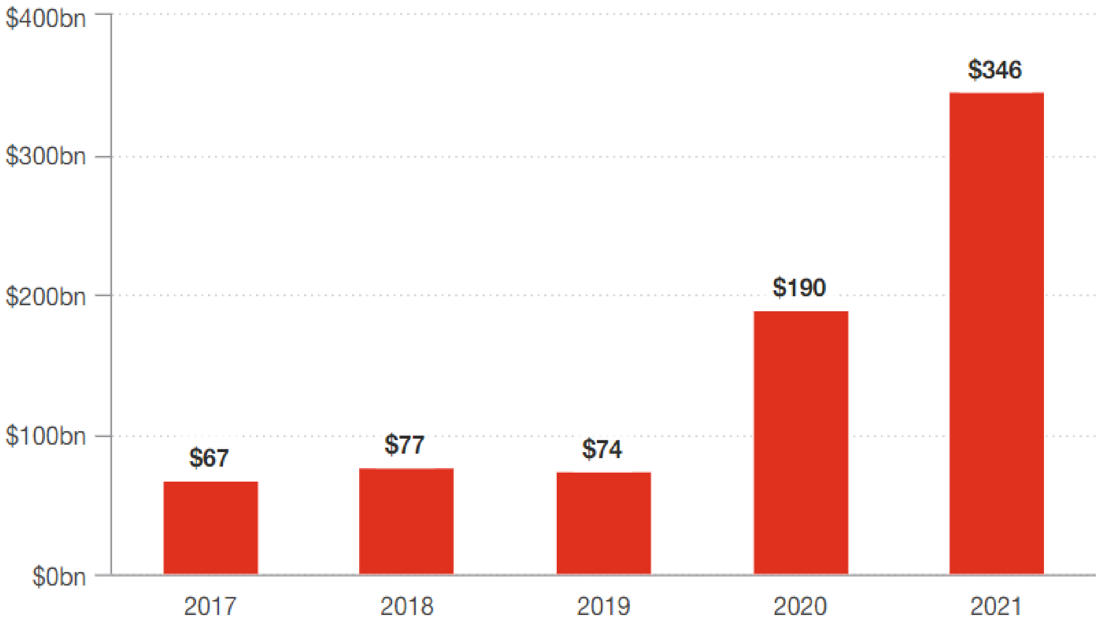
The 1st-day change in price after the IPO date

5-year Treasury bill

IPO Market

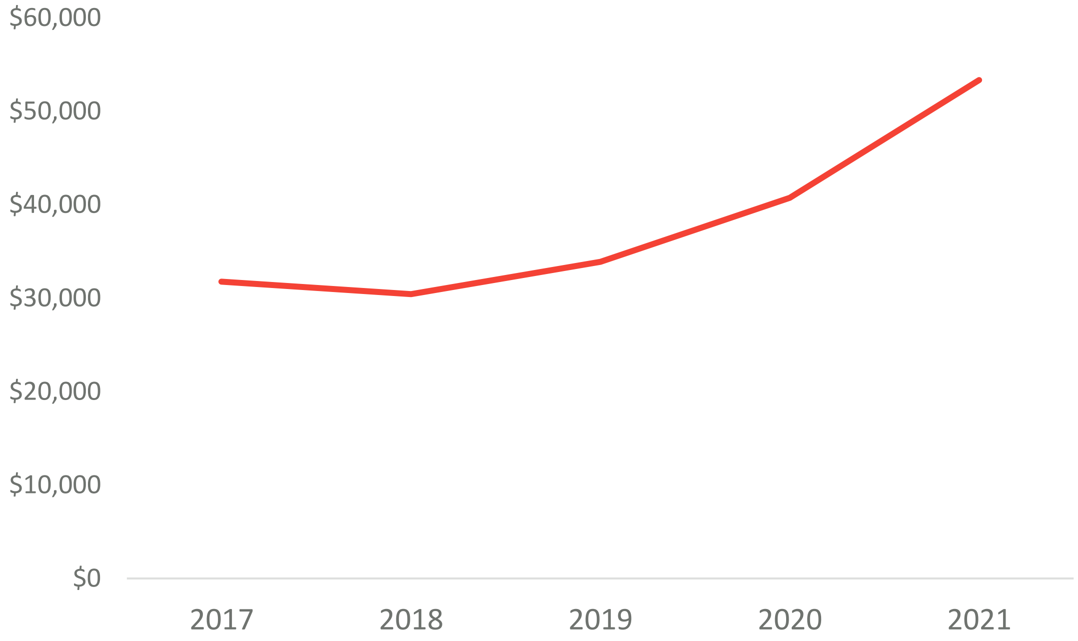
2021 was the biggest IPO year ever - extraordinary volumes globally with $608bn raised

Americas IPO proceeds ($bn)



Source: Dealogic

U.S. Equity Market Value ($bn)

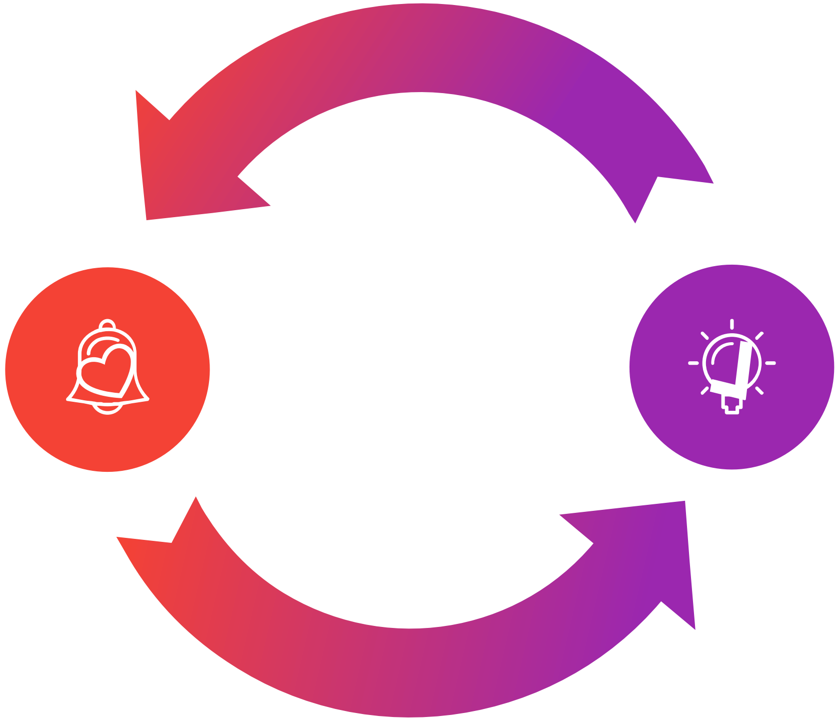


Main Challenges

How to build an accurate model for assessing new IPO investments
without quantifying issuers' non-financial factors and underwriters' conflict of interest?

Non-financial factors

Investors base an average of 40% of their IPO investment decisions on non-financial factors, especially quality of management, corporate strategy and execution, brand strength and operational effectiveness, a compelling equity story and corporate governance (according to an EY report).



Underwriters' conflict of interest

The success or failure of an IPO is greatly determined by an accurate pricing. Valuation and pricing are complex processes and are hugely important components of the IPO process. Given the fact that often there are divergent interests of the issuer and the underwriter, the valuation and pricing do not necessarily go hand in hand.

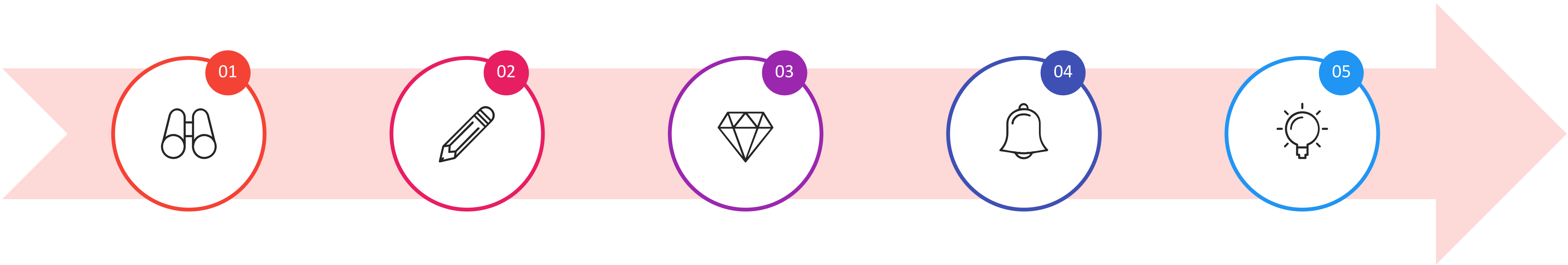
Solution: HyIPO

Tool for picking good companies intending to IPO

GOAL

Risk-return assessment for upcoming IPO investments

Methodology



Data acquisition and understanding

Data analysis and features selection

Building the models

Conclusions and model interpretability

Limitations and future recommendations

Data Acquisition

Dataset provided by the IPOScoop website includes information about the **Issuer, Symbol, Rating, IPO date, IPO price, the 1st-day returns and the IPO managers** for the period 2000-2020 (3633 observations)

The Rating *is a consensus taken from Wall Street and investment professionals concerning how well an IPO might perform when it starts trading*

The **target variable** is "label 1" if the 1st-day change in price after the IPO date is higher than the risk-free rate benchmark (5-year Treasury bill rate) and "label 0" otherwise

Data cleaning: standard procedures (changing data types, checking whether NaN's and/or Null values exist, dropping useless columns, etc.) plus checking and replacing the issuers' trading symbols in the IPOs list with the accurate ones from a dataset which captures the US Publicly Listed Companies as of today

Data Acquisition

Feature engineering and extraction

- the 1st-week and 1st-month closing prices subsequent to the IPO (Yahoo Finance)
- **market performance indicator**: the change in S&P500 closing prices for 1 week, 1 month and 3 months prior to the IPO date (Yahoo Finance)
- **market volatility indicator:** the VIX change in closing values for 1 week, 1 month and 3 months prior to the IPO date (Yahoo Finance)

(the CBOE Volatility Index, or VIX, is a real-time market index representing the market's expectations for volatility over the coming 30 days)

- **AAII Investor Sentiment Survey (bull-bear spread)** published during the week prior to the IPO date

(the participants in the survey answer the following question: what direction do AAII members feel the stock market will be in the next 6 months?)

- "label 1" if the Lead/Joint-Lead Managers are Tier 1 underwriters or "label 0" otherwise
- **date-based features** (day of the week)
- **social indicator:** search data from google trends API in order to assert potential investors' appetite for each IPO; **the number of spikes in reported popularity** (if observation > mean) during the last 2 weeks prior to the IPO
- 5-year Treasury bill historical rate for each IPO date

EDA & Features Selection

Feature engineering and extraction

Initial set of features: '1st Day % Px Chng ', 'Star Ratings', 'S&P 1 Week % Px Chng', 'S&P 1 Month % Px Chng', 'S&P 3 Months % Px Chng', 'VIX 1 Week % Px Chng', 'VIX 1 Month % Px Chng', 'VIX 3 Months % Px Chng', 'Sentiment_survey', 'Top IB', 'weekday

Performed correlation analysis of the features

Applied ML models and ANN with different sets of features and then examined the features' importance and the metrics

Unnecessary or correlated features decrease training speed, model interpretability and the generalization performance on the test set

Selected set of features: 'Star Ratings', 'S&P 3 Months % Px Chng', 'VIX 1 Week % Px Chng', 'Sentiment survey', 'Tier1 IB'

Models

Binary classification problem

- split data into train/ test sets
- define a pipeline for preprocessing
- use PyCaret to determine the 5 best model in terms of accuracy
- create a Custom Metric in PyCaret - Profit - to select the model which maximizes the business value: if we predict investing in the IPO and the real value is 1, we gain the potential profit (based on historical average) * investment value; if we predict investing in the IPO and the real value is 0, we take the potential loss (based on historical average) * investment value
- fine-tune the hyperparameters of the best models using Grid Search
- evaluate the performance of the models on the test dataset
- build NN architectures and experiment with more aspects of Dense NN models such as layer activations, learning rates, regularization
- select the best model in terms of accuracy, f1-score, business value and explainability – Logistic Regression

Models

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	Profit	TT (Sec)
rbfsvm	SVM - Radial Kernel	0.7473	0.7688	0.7490	0.7976	0.7723	0.4890	0.4904	4114.9245	0.466
lr	Logistic Regression	0.7462	0.7864	0.7470	0.7972	0.7711	0.4868	0.4883	4103.6007	0.050
ada	Ada Boost Classifier	0.7387	0.7731	0.7440	0.7882	0.7650	0.4713	0.4729	4079.1372	0.224
gbc	Gradient Boosting Classifier	0.7302	0.7740	0.7450	0.7754	0.7596	0.4523	0.4532	4072.8346	0.290
knn	K Neighbors Classifier	0.6914	0.7319	0.7430	0.7250	0.7338	0.3667	0.3670	4012.1574	0.194
svm	SVM - Linear Kernel	0.7410	0.0000	0.7221	0.8064	0.7611	0.4801	0.4845	3974.0176	0.032
lightgbm	Light Gradient Boosting Machine	0.6960	0.7593	0.7251	0.7397	0.7320	0.3806	0.3812	3929.6854	0.134
nb	Naive Bayes	0.7325	0.7769	0.7101	0.8006	0.7524	0.4636	0.4678	3903.8315	0.038
et	Extra Trees Classifier	0.6868	0.7071	0.7211	0.7296	0.7250	0.3612	0.3616	3898.0645	0.708
rf	Random Forest Classifier	0.6880	0.7434	0.7191	0.7319	0.7251	0.3642	0.3647	3889.7318	0.776
dt	Decision Tree Classifier	0.6355	0.6325	0.6544	0.6926	0.6726	0.2621	0.2630	3501.5181	0.034

Conclusions

Model interpretability (Logistic Regression)

- The IPO rating (1 to 5 hierarchical values) has a positive impact on the odds that the IPO returns represented in the observation are in the target class ("1")
- The market performance indicator - the change in S&P500 closing prices for 3 months prior to the IPO date has a positive impact on the odds that the IPO returns represented in the observation are in the target class ("1")
- The forward-looking AAII Investor Sentiment Survey (bull-bear spread), published during the week prior the IPO date has a positive impact on the odds that the IPO returns represented in the observation are in the target class ("1")
- The Lead/Joint-Lead Managers being Tier 1 underwriters date has a slightly positive impact on the odds that the IPO returns represented in the observation are in the target class ("1")
- The market volatility indicator - the change in VIX closing values for 1 week before the IPO has a positive impact on the odds that the IPO returns represented in the observation are NOT in the target class ("1")

	coef
Star Ratings	3.945052
S&P 3 Months % Px Chng	2.598894
Sentiment_survey	1.790769
Top IB	1.088593
VIX 1 Week % Px Chng	0.532425

Conclusions

Model interpretability (Logistic Regression)

```
Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.80      0.71       276
           1       0.80      0.64      0.71       351

    accuracy                           0.71       627
   macro avg       0.72      0.72      0.71       627
weighted avg       0.73      0.71      0.71       627

-----------------------------------------------------------
Standard Confussion Matrix (error matrix):
 [[220  56]
 [125 226]]
-----------------------------------------------------------
Accuracy Score obtained is: 71.13%
-----------------------------------------------------------
f1_macro Score obtained is: 71.13%
-----------------------------------------------------------
f1_micro Score obtained is: 71.13%
-----------------------------------------------------------
f1_weighted Score obtained is: 71.16%
-----------------------------------------------------------
f1 Score obtained is: 70.85%
```

Limitations

Future Recommendations

Parse the S-1 and F-1 Reports

limitation: after conducting web scrapping with the requests HTML package and retrieving the URLs for the S-1 and F-1 reports from SEC website, some of the reports were missing and some were incorrectly selected

future recommendation: improve the web scrapping technique to properly parse the S-1 Reports and perform a sentiment analysis, extract financial factors, particularly debt to equity ratios, EPS growth, sales growth, ROE, profitability and EBITDA growth and examine the risk factors section of the reports

Perform sentiment analysis on companies' news

limitation: when analyzing social indicators - data from google trends API - more specifically the number of spikes in popularity registered by the issuer during the last 2 weeks before the IPO, we cannot state if the spikes are due to good or bad news

future recommendation: develop a framework for automatically distilling stock market insights from an online conversation (news articles, social media posts, etc.) before the IPO date and perform sentiment analysis

Peer benchmarking

future recommendation: build a benchmark with competitors and comparable companies for each issuer in order to assess the analyzed company's business fundamentals with its peers group

Thank you