Classification of start-ups using supervised Learning

Carolina Gonzalez

Business Case

- The model evaluates the classification of start-ups between:
  - Acquired
  - IPO
  - Other

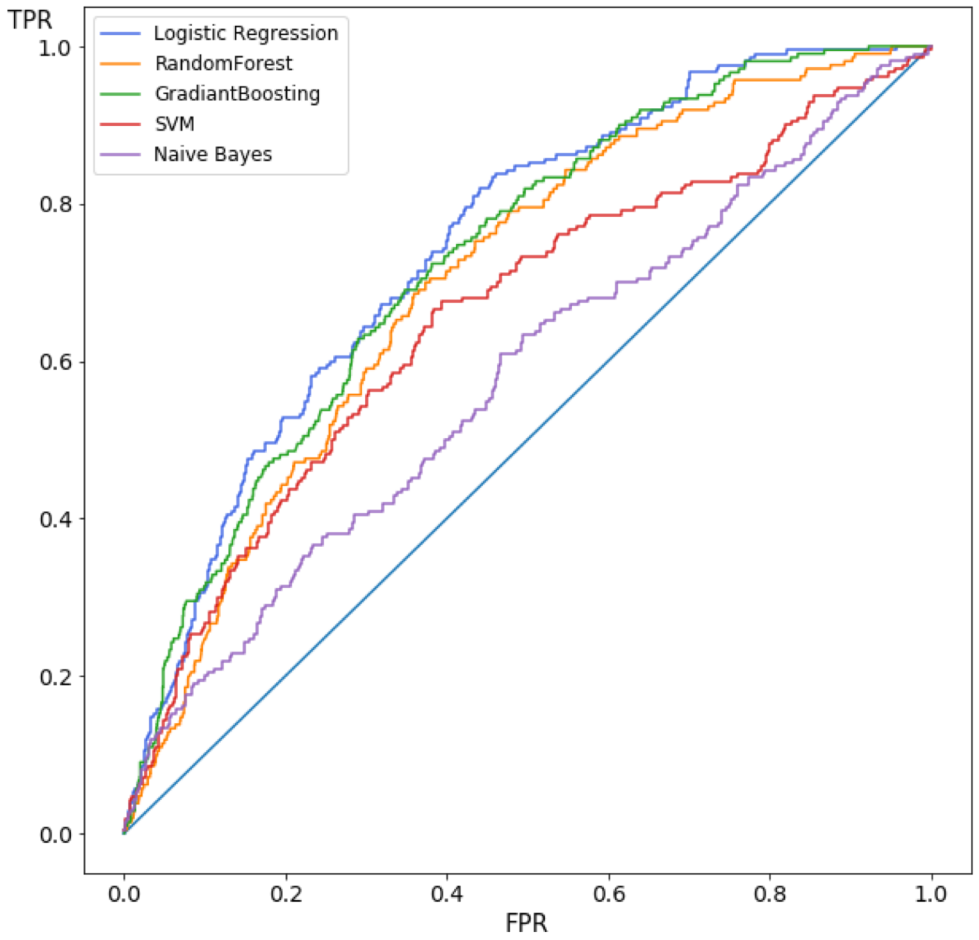
- This model could help:
  - Investors: prioritize start-ups for funding evaluation
  - Start-up employees: evaluate the likelihood of shares maturing
  - Entrepreneurs: review of potential exit strategies

The Data

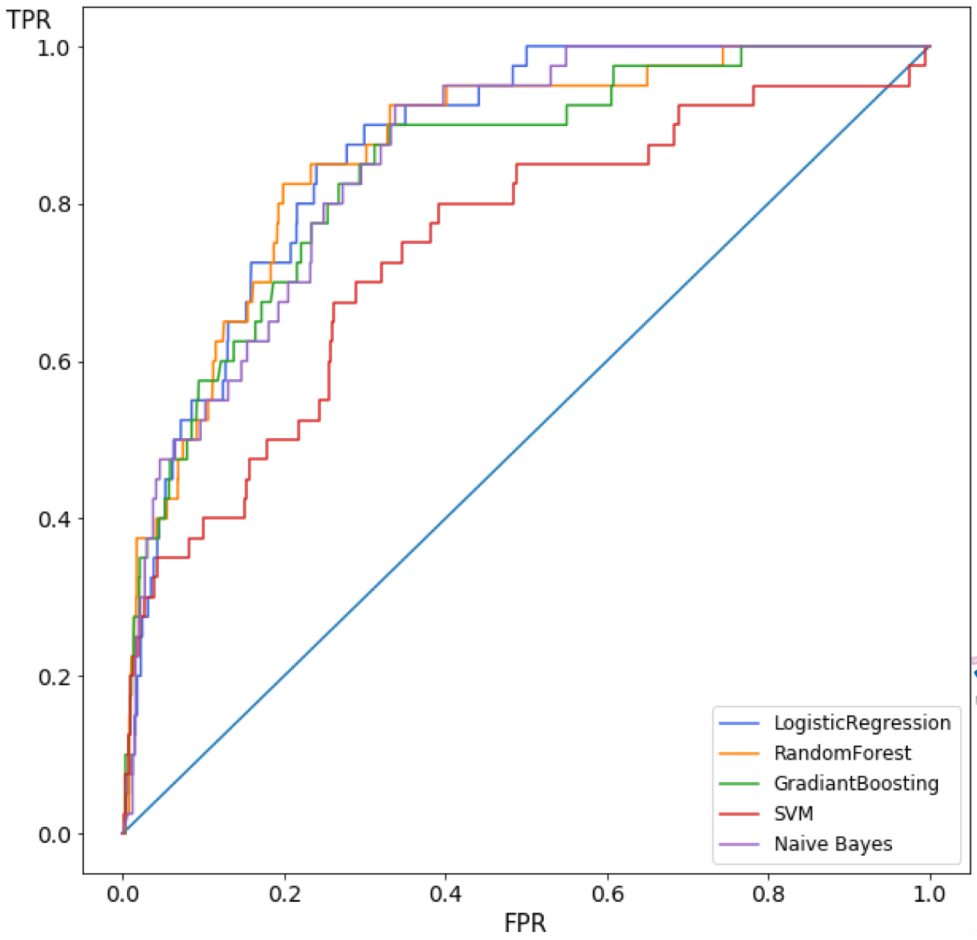
- Published CPI by US Bureau of Labor Statistics
- Data collected by Crunchbase
- Data up to 2013 for start-ups in the US
- Considered companies 5+ years old
- Acquired 12.3%, IPO 2.4%
- Features considered:
  - Average annual inflation the year the company was founded
  - Month founded
  - Company category
  - State founded
  - Inflation adjusted raised amounts from crowdfunding and angel investors
  - Time (days) in operation at the time first founding was raised

Model Selection

ROC curves for evaluated models for Acquisitions



ROC curves for evaluated models for IPOs



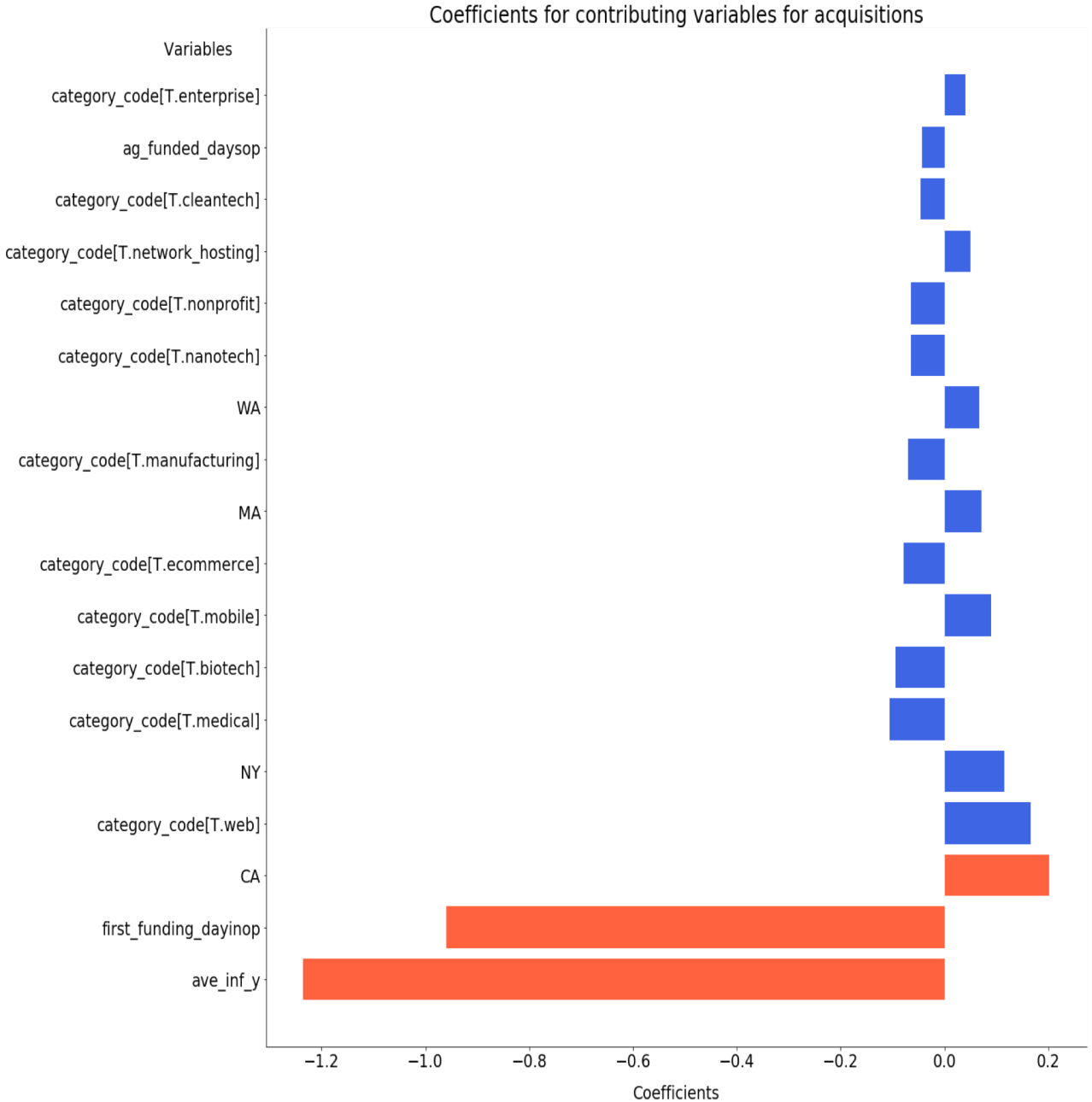
The model selected

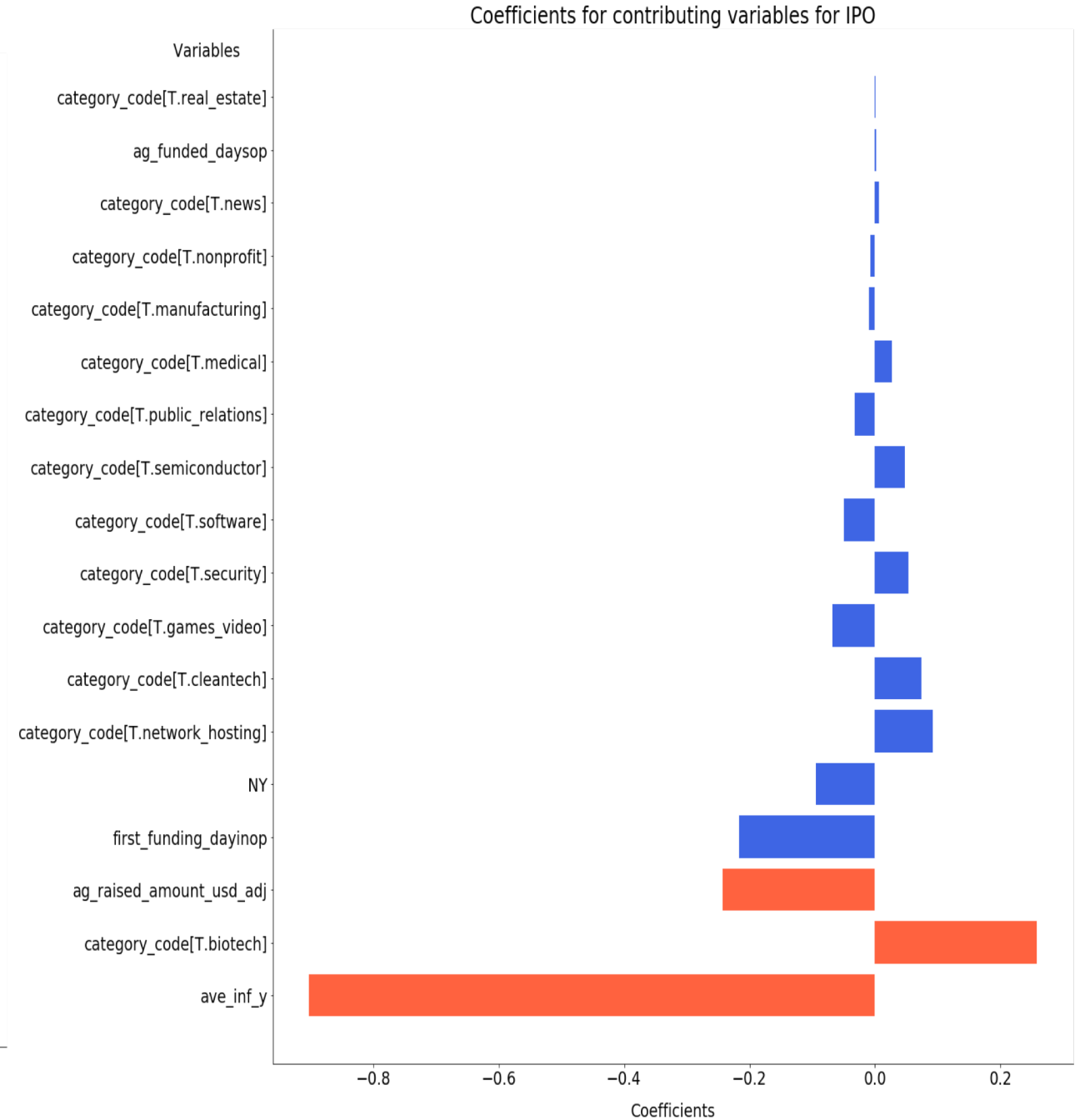
- Logistic regression
  - High AUC
  - Interpretable
  - Scalable
  - Not overfitted
- Lasso regularization, C = 0.1
- Threshold selected to maximize Precision

	Acquired	IPO
Auc_test	0.74	0.87
Auc_train	0.76	0.88
Threshold	0.46	0.34
Precision, avg.	0.84	0.96

Acquired	Classified Negative	Classified Positive
True negatives	1392	8
True positives	204	6

IPO	Classified Negative	Classified Positive
True negatives	1563	7
True positives	38	2


Coefficients for contributing variables for acquisitions
Variables
category_code[T.enterprise]
ag_funded_daysop
category_code[T.cleantech]
category_code[T.network_hosting]
category_code[T.nonprofit]
category_code[T.nanotech]
WA
category_code[T.manufacturing]
MA
category_code[T.ecommerce]
category_code[T.mobile]
category_code[T.biotech]
category_code[T.medical]
NY
category_code[T.web]
CA
first_funding_dayinop
ave_inf_y
−1.2
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
Coefficients


Coefficients for contributing variables for IPO
Variables
category_code[T.real_estate]
ag_funded_daysop
category_code[T.news]
category_code[T.nonprofit]
category_code[T.manufacturing]
category_code[T.medical]
category_code[T.public_relations]
category_code[T.semiconductor]
category_code[T.software]
category_code[T.security]
category_code[T.games_video]
category_code[T.cleantech]
category_code[T.network_hosting]
NY
first_funding_dayinop
ag_raised_amount_usd_adj
category_code[T.biotech]
ave_inf_y
−0.8
−0.6
−0.4
−0.2
0.0
0.2
Coefficients

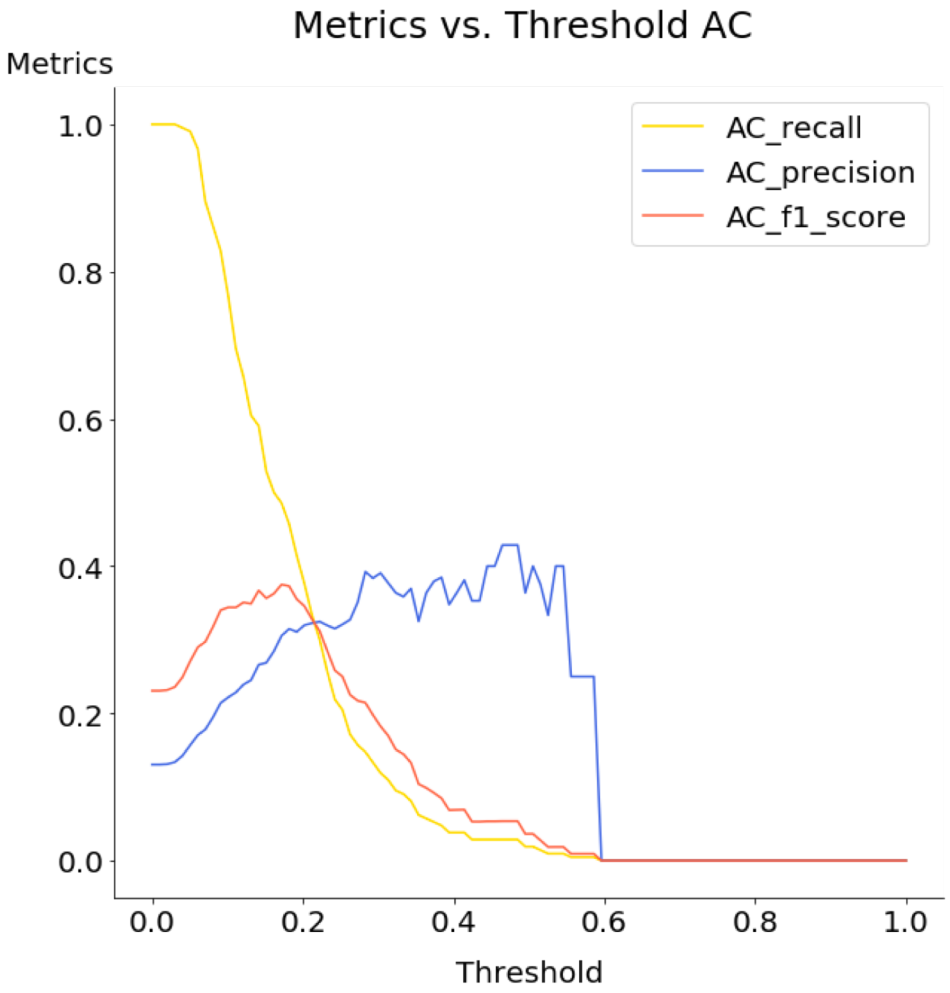
Conclusion

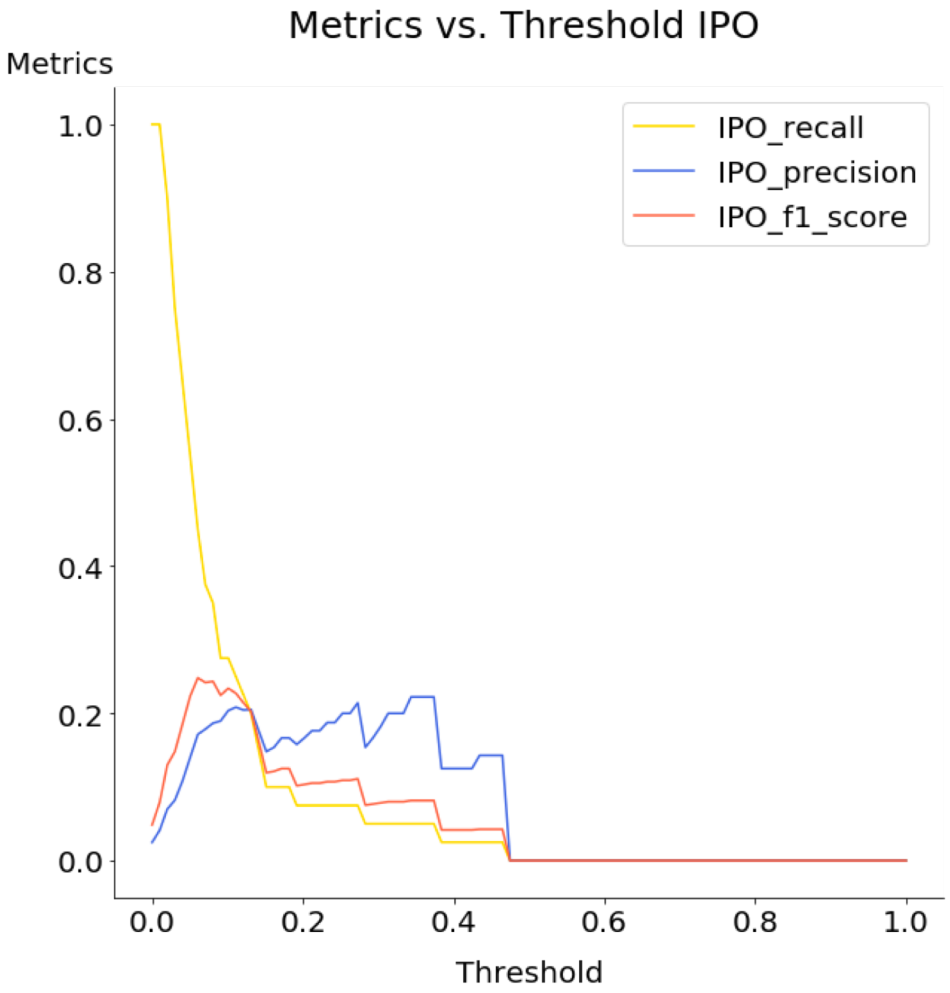
- A logistic model was developed to classify start-ups 5+ years old into:
  - Acquired
  - IPO
- Insights for main contributing variables were extracted.
- Future work:
  - Include more data from recent years
  - Look at the effect of crowdfunding on the evolution of companies.
  - More data for IPO companies

Questions ?

Appendix

Threshold evaluation





Model Selection

Acquisition Models

Model	Auc_test	Auc_train
Log Regression Normalized	0.74	0.76
Random Forest	0.70	0.78
Gradient Boosting	0.73	0.82
Support Vector Machine	0.66	0.68
Naïve Bayes	0.58	0.60

IPO Models

Model	Auc_test	Auc_train
Log Regression Normalized	0.87	0.88
Random Forest	0.87	0.90
Gradient Boosting	0.85	0.92
Support Vector Machine	0.74	0.82
Naïve Bayes	0.86	0.83

Metrics IPO: investing use case

Metrics	Precision	Recall	F1-score	Support
0	0.98	1.0	0.99	1570
1	0.22	0.05	0.08	40
Avg./Total	0.96	0.97	0.96	1610

Categories	Classified Negative	Classified Positive
True negatives	1563	7
True positives	38	2

Threshold= 0.34

Metrics AC: investing use case

Metrics	Precision	Recall	F1-score	Support
0	0.92	0.82	0.87	1400
1	0.31	0.49	0.38	210
Avg./Total	0.84	0.79	0.81	1610

Categories	Classified Negative	Classified Positive
True negatives	1392	8
True positives	204	6

Threshold= 0.46

Metrics AC: filtering use case

Metrics	Precision	Recall	F1-score	Support
0	0.92	0.82	0.87	1400
1	0.31	0.49	0.38	210
Avg./Total	0.84	0.79	0.81	1610

Categories	Classified Negative	Classified Positive
True negatives	1168	232
True positives	108	102

Threshold= 0.17

Metrics IPO: filtering use case

Metrics	Precision	Recall	F1-score	Support
0	0.99	0.94	0.96	1570
1	0.17	0.45	0.25	40
Avg./Total	0.97	0.93	0.95	1610

Categories	Classified Negative	Classified Positive
True negatives	1483	87
True positives	22	18

Threshold= 0.06