Optimizing Startup Investments

Daniel Rapoport, Maxime Basse, Aziz Malouche

Problem Description

Predicting Start-Up Success is hard…

Can we yield better predictions with techniques learned in class?

Problem Description

Predicting Start-Up Success is hard…

Can we yield better predictions with techniques learned in class?

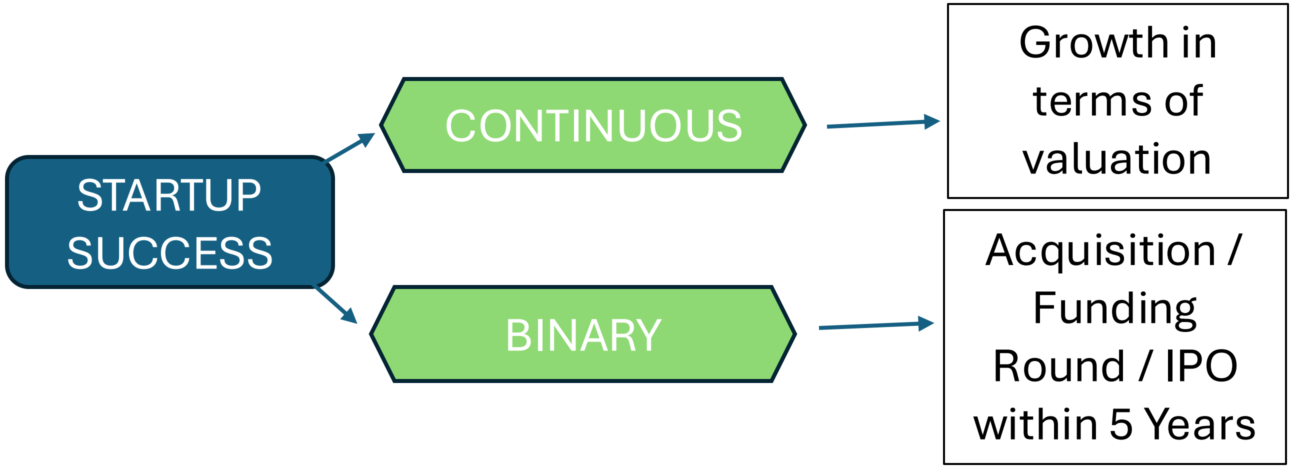
How can robust approaches improve our investment decisions?

Data and Defining Startup Success

Dataset from Crunchbase

- 10,000 startups
- 22,000 funding rounds
- 3,200 acquisitions

Success Assessment
How does a startup perform in the 5 years following its Series B?



Significant Features
Earliest Funding Round – Type
Earliest Funding Round - Number of Funding Rounds
Days Between Founded and Earliest Funding Round
Earliest Funding Round - Money Raised (in USD)
Series A Money Raised (in USD)
Total Money Raised Before Series B
Days Between Founded and Series B
Total Funding Rounds Before Series B
Number of Seed/Pre-Seed Rounds
Days Between Founding and First Seed/Pre-Seed
Days Between Founding and Last Seed/Pre-Seed
Days Between Founding and Series A

Results – Traditional Predictions

Model	Numerical	Numerical + Categorical	Numerical + TabText (Cat.)	Numerical + Cat. + Bert (Description)	Numerical + Cat. + Llama (Desc.)	Numerical + Cat. + Transfer-Learning (Llama)
CART	0.61	0.66	0.61	0.61	0.61	0.63
Random Forest	0.75	0.78	0.73	0.72	0.72	0.76
XG-Boost	0.75	0.79	0.77	0.75	0.75	0.77
LightGBM	0.76	0.79	0.78	0.77	0.76	0.79
CatBoost	0.78	0.80	0.79	0.78	0.77	0.80

Methods – Predictions to Prescriptions

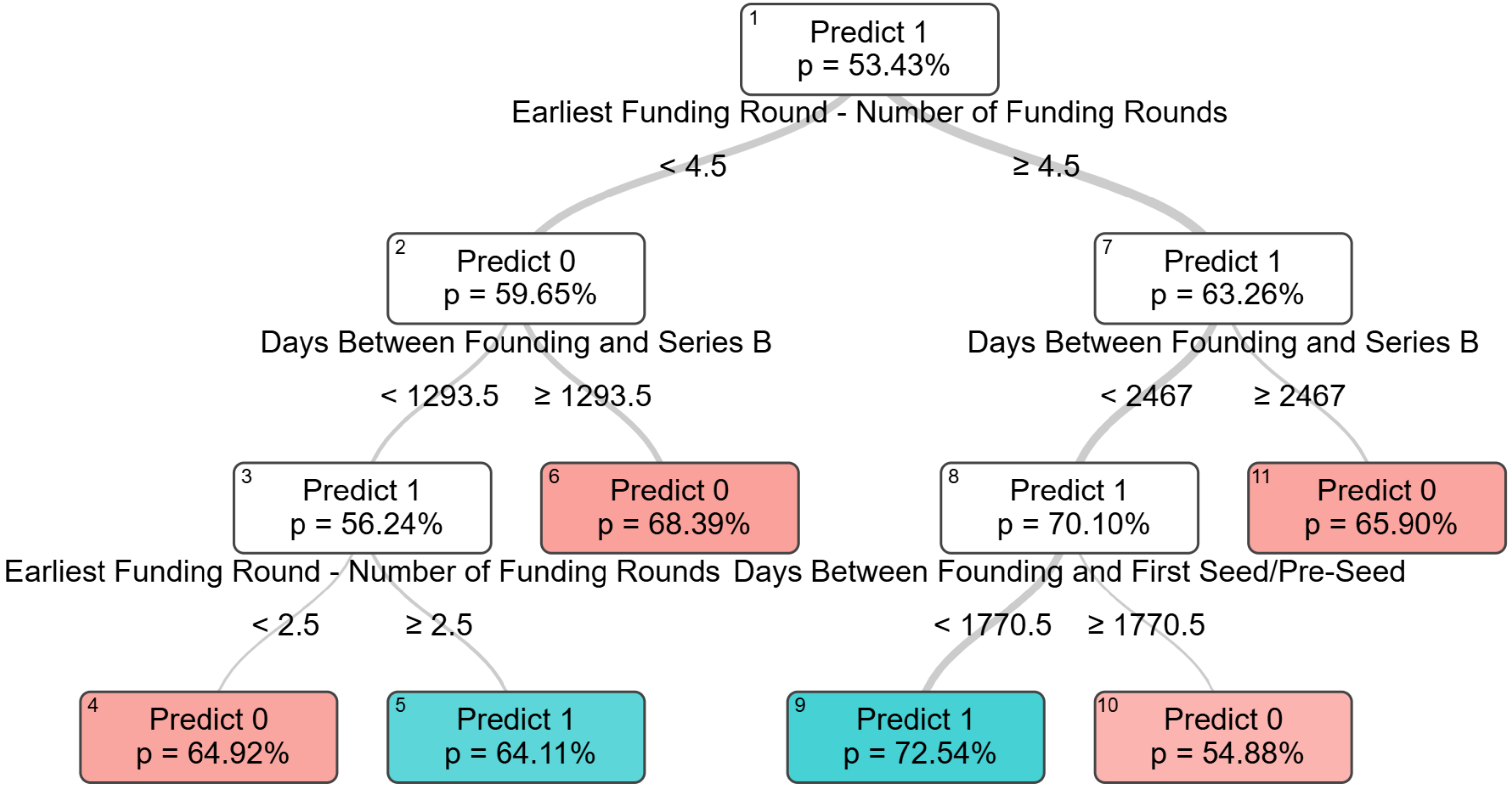
Point Prediction → **No uncertainty** assumption of prediction

Robust Prediction → Assume **general uncertainty of** predictions

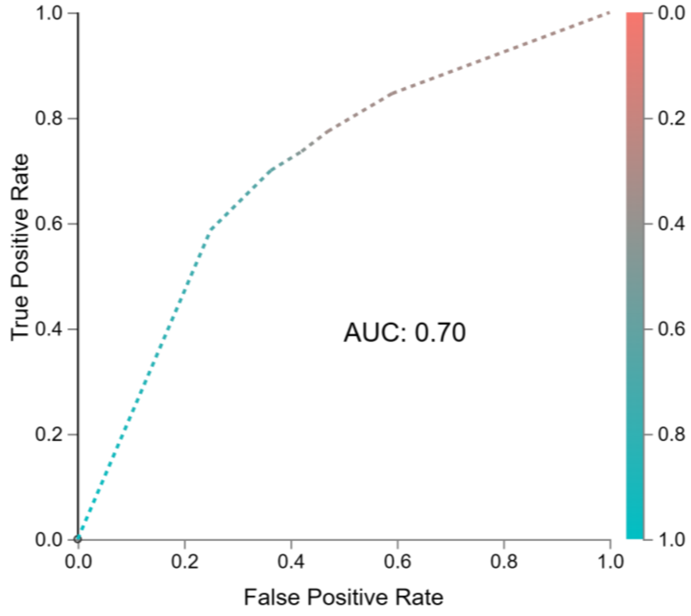
Weighted Average → **Neighbor's** of sample might **indicate uncertainty** of it

Results - OCT

Binary Optimal Classification Tree



ROC Curve



Methods – Predictions to Prescriptions

Point Prediction

$$\max \sum_{i=1}^{N} \hat{y}(x^i)\hat{r}(x^i)z_i, \quad \text{s.t.} \quad \underbrace{\sum_{i=1}^{N} c_i z_i \leq B, \quad z_i \geq 0}_{\text{Investment Budget}} \quad \forall i.$$

Robust Point Prediction

$$\underbrace{\mathcal{U} = \left\{ \Delta y \in \{0,1\}^N \,\middle|\, \frac{1}{N}\sum_{i=1}^{N} \Delta y_i \leq \Gamma \right\}}_{\text{Uncertainty set}}.$$

$$\max_{z} \min_{\Delta y \in \mathcal{U}} \sum_{i=1}^{N} |\hat{y}(x^i) - \Delta y_i| \hat{r}(x^i) z_i,$$

$$\text{s.t.} \quad \sum_{i=1}^{N} c_i z_i \leq B, \quad z_i \geq 0, \quad \Delta y_i \in \{0,1\}, \quad \forall i.$$

Weighted Average

$$\max \sum_{i=1}^{N} \left(\overbrace{\frac{1}{|L_1(s_i) \cup L_2(s_i)|} \sum_{u_j \in L_1(s_i) \cup L_2(s_i)}}^{\text{Leaf Neighbor Average}} r(u_j) y(u_j) \right) z_i$$

$$\text{s.t.} \quad \sum_{i=1}^{N} z_i \leq B, \quad z_i \geq 0 \quad \forall i$$

Results– Predictions to Prescriptions

	Median Return			Risk-Adjusted Return		
K (= size of portfolio)	**Point** Prediction	**Weighted** Average	**Robust** Point Pred.	**Point** Prediction	**Weighted** Average	**Robust** Point Pred.
10	9.6	**11.5**	5.8	0.27	**0.29**	0.24
25	15.7	**15.9**	9.4	0.30	0.29	**0.43**
50	16.6	**18.2**	17.2	0.36	0.37	**0.97**
100	3	20.8	**23.7**	0.29	0.34	**1.13**

**Optimization problem solved for test-set of size n with k start-ups considered for each of the portfolios. Displays realized return with budget of $1000 in thousands.*

Discussion and Conclusion

- Start-up data poses challenges regarding data availability
- Defining a clear metric of success is critical
- Unstructured features lack predictive value compared to numerical and categorical features
- Weighted Average achieves the highest median return
- Robust point prediction yields the highest Sharpe Ratio