Predicting Series A Achievement for Pre-Seed Startups: A Machine Learning Approach

Makhmud Islamov Sukhrobjon Golibboev

Department of Computer Science, Stanford

Predicting

The venture capital landscape faces a critical challenge: approximately 90% of startups fail. We developed a machine learning framework to predict whether pre-seed startups will achieve Series A funding or acquisition. Our algorithm processes multidimensional startup data including founding team characteristics, economic indicators, and venture capital market conditions to output a binary prediction.

Data Source & Composition

- Source: partnership with Aviato for proprietary startup data via API
- Dataset: U.S.-based startups from 2012-2022 period, initially ~170,000 records.
- Final dataset: ~120,000 records, 66 columns after adding new features
- Class distribution: 15.85% successful (reached Series A or acquired)

Feature Engineering

Industry Classification (LDA)

- Used Latent Dirichlet Allocation to cluster startups into 15 industry sectors
- Facilitates normalization of industry-specific funding patterns
- Optimized using coherence score with 5-fold cross-validation

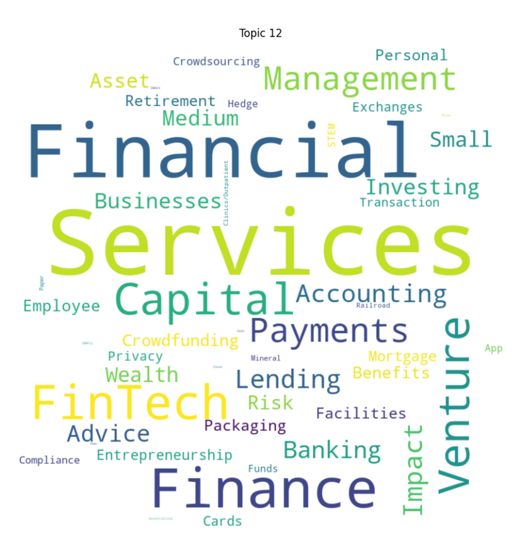

Figure 1. LDA industry clusters example

Economic & VC Market Indicators

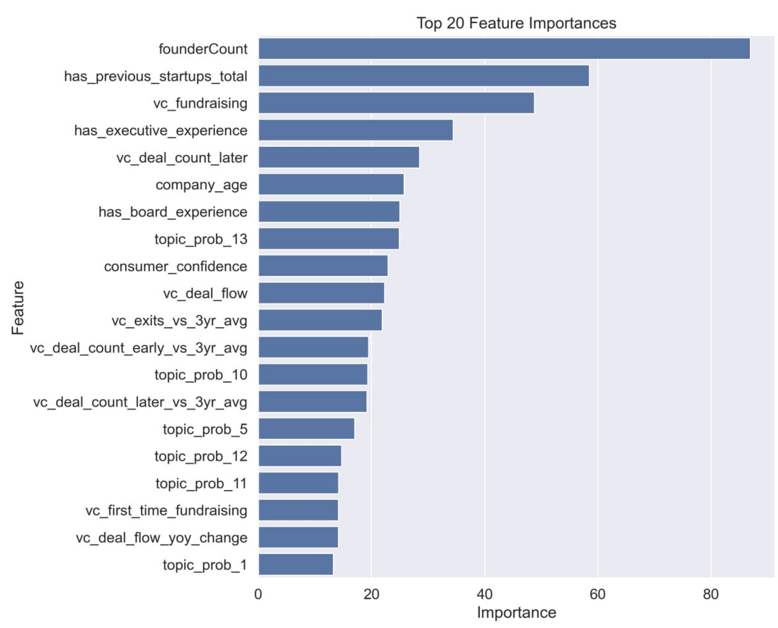

Figure 2. Feature importance ranking showing dominance of founder-related attributes

Economic Indicators

- **GDP Growth Rate:** Correlates with investment appetite
- **Consumer Confidence:** Indicates potential consumer adoption
- **Treasury Yield Curve:** Predicts economic cycles affecting VC readiness

VC Market Indicators

- **Deal Values:** Shows current VC activity level and momentum
- **Exit Counts:** Liquidity options influencing investment decisions
- **Fundraising Totals:** Demonstrates LPs' confidence in venture capital

Team Competency Features

Feature (Range)	Description
Team Industry Alignment (0-1)	Congruence between team's experience and startup sector
Team Expertise Diversity (0-1)	Complementarity of expertise across founding members
Competency Compound Score (0-1)	Weighted combination (0.7 × alignment + 0.3 × diversity) balancing domain-specific expertise with skill complementarity

Table 1. Quantitative metrics to assess founding team capabilities

Data & Model Pipeline

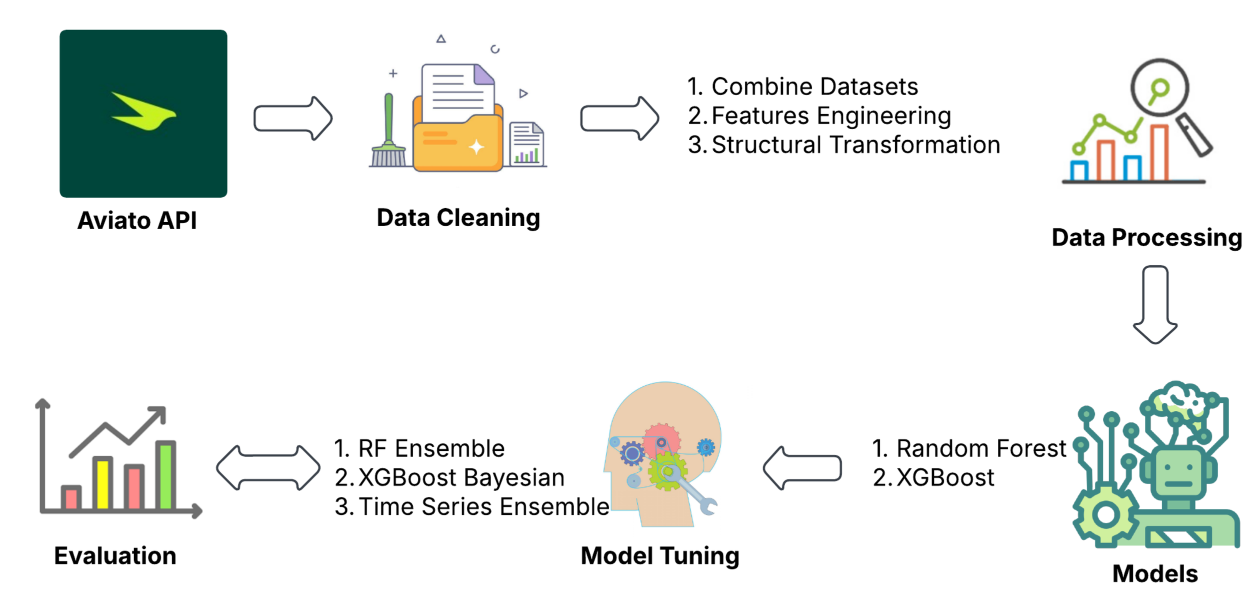

Figure 3. Machine Learning Pipeline Diagram

Model Implementation

1. **Random Forest Ensemble:**
   - Multiple RF models with varied configurations
   - n_estimators=500, class_weight={0:1, 1:5}
   - Decision threshold optimization (0.35)
2. **XGBoost with Bayesian Optimization:**
   - Hyperparameters: max_depth=6, learning_rate=0.03138
   - subsample=0.8787, colsample_bytree=0.5756
   - min_child_weight=10, scale_pos_weight=4.6924
3. **Time-Specific Models:**
   - Separate models for different periods (early/mid/recent)
   - Time interaction features (e.g., founderCount_x_period)

Mathematical Formulation

- **Random Forest:** $\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$
- **Ensemble Decision:** $\hat{y} = \begin{cases} 1, & \text{if } \frac{1}{n}\sum_{i=1}^{n} p_i(x) \geq \theta \\ 0, & \text{otherwise} \end{cases}$
- **XGBoost Objective:** $L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$

Feature Importance & Analysis

Key Feature Findings

- **Founder characteristics** emerge as the strongest predictors of startup success, confirming prior research by Gompers et al. (2010)
- **VC ecosystem indicators** provide crucial temporal context for funding likelihood
- **Industry sectors** (topic_prob features) demonstrate sector-specific success patterns
- **Time period matters** - feature importance shifts across different time periods

Results

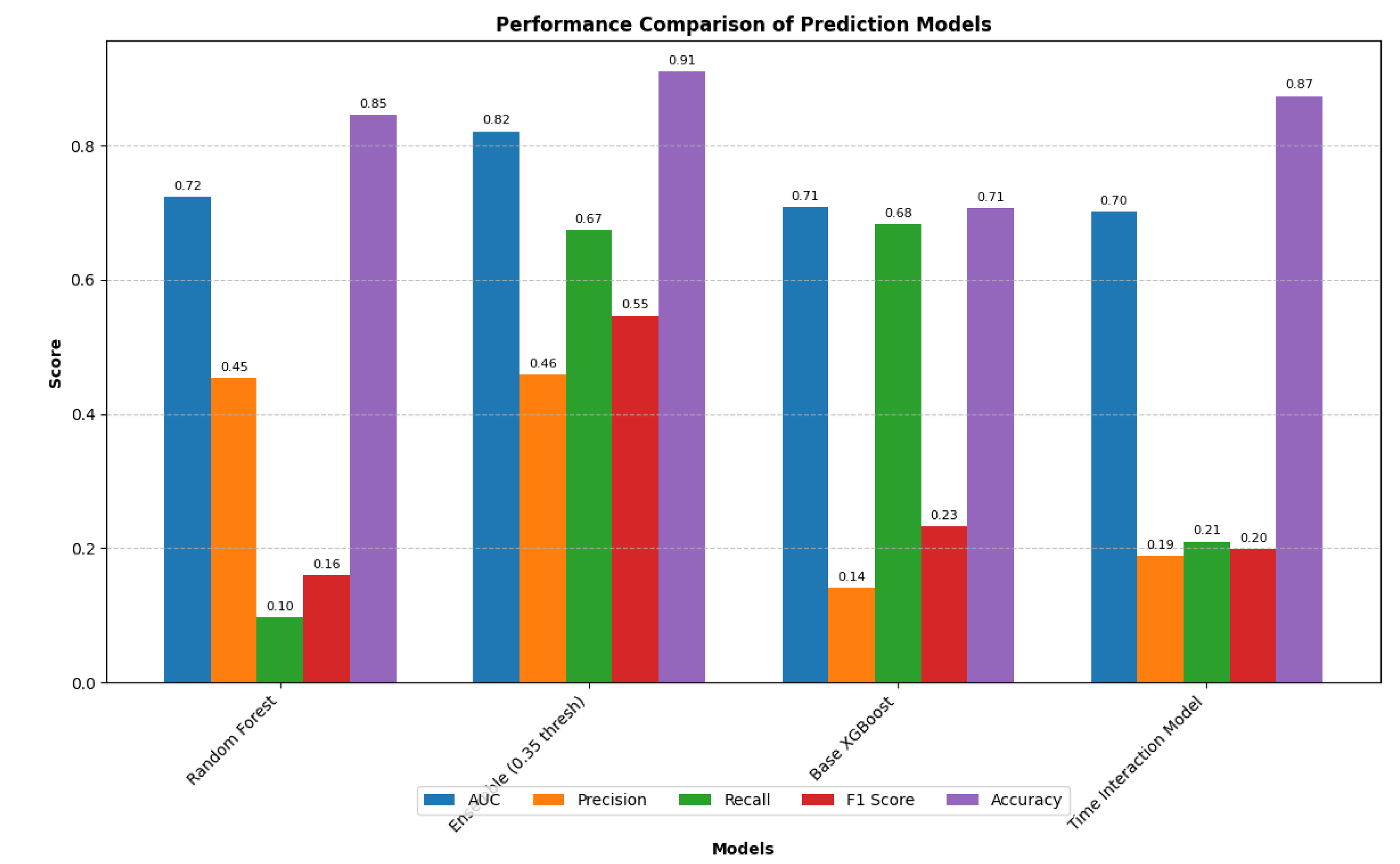

Figure 4. Performance scores comparison across models showing superior performance of the Ensemble approach

Discussion

Model Performance Insights

- The ensemble model significantly outperformed individual models across key metrics
- Optimizing the classification threshold (0.35) provided the best precision-recall balance
- Temporal drift in startup success patterns presents a fundamental challenge for predictive modeling
- Time-specific models show promise but performance on recent startups remains challenging

Future Work

- **Refine success definition** with clear funding and time-bound objectives (e.g., Series A within four years, $8M+ funding)
- **Expand dataset** temporally (1999-2022) and geographically, to new startup hubs globally
- **Enhance model robustness** through increased sample size and ensemble methodologies
- **Address temporal drift** with more sophisticated time-adaptive modeling approaches

References

1. Gompers, P., Kovner, A., Lerner, J., & Scharfstein, D. (2010). Performance persistence in entrepreneurship. Journal of Financial Economics, 96(1):18-32.
2. Ang, Y.Q., Chia, A., & Saghafian, S. (2022). Using machine learning to demystify startups' funding, post-money valuation, and success. Springer International Publishing, pages 271-296.
3. Corea, F., Bertinetti, G., & Cervellati, E.M. (2021). Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors. Machine Learning with Applications, 5:100062.
4. Li, Y., Zadehnoori, I., Jowhar, A., Wise, S., Laplume, A., & Zihayat, M. (2024). Learning from yesterday: Predicting early-stage startup success for accelerators through content and cohort dynamics. Journal of Business Venturing Insights, 22:e00490.
5. Razaghzadeh Bidgoli, M., Raeesi Vanani, I., & Goodarzi, M. (2024). Predicting the success of startups using a machine learning approach. Journal of Innovation and Entrepreneurship, 13(1):80.

 makhmud@stanford.edu, sukhrob@stanford.edu