Founder Matching: Optimizing Team Formation with Machine Learning

INDENG 242 – Machine Learning and Data Analytics

Idris Houir Alami
Youssef Miled
Ryan Michael Chekkouri

Motivation

- Team quality is a key determinant of startup success
- Founder matching is often informal and heuristic-based
- Existing platforms lack data-driven compatibility modeling

Goal: Build a scalable, interpretable ML system for cofounder matching that balances

- similarity (vision, communication style)
- complementarity (skills, roles, experience)

Dataset

- Synthetic but realistic dataset of $\sim$ 1,200 founders
- Generated using LLM-based persona construction

Feature types:

- Categorical: industry, preferred role
- Multi-label: tech stack, strengths, weaknesses, roles
- Numerical: risk tolerance, collaboration openness, experience
- Text (idea title/description): used for similarity analysis

Learning Problem

- Founders act as both "users" and "items"
- Objective: predict compatibility score R_{ij} between founders i and j

$$R_{ij} = \alpha S_{ij} + (1 - \alpha) C_{ij}, \quad \alpha = 0.5$$

- S_{ij}: similarity score
- C_{ij}: complementarity score

Feature Engineering

Similarity Features

- Industry
- Behavioral scores (risk, communication, responsiveness)
- Idea semantics

Complementarity Features

- Roles and preferred roles
- Tech stack
- Strengths and weaknesses
- Experience and education

Similarity Computation

Similarity is computed using cosine similarity:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

- Applied to L2-normalized similarity feature matrix
- Captures alignment in vision, style, and industry

Complementarity Computation

- Multi-label features use Jaccard distance:

$$C_{ij} = 1 - \frac{|A \cap B|}{|A \cup B|}$$

- Ignores co-absence
- Emphasizes functional diversity
- Numerical features use standardized distances

Collaborative Filtering Model

We apply matrix factorization:

$$\hat{R}_{ij} = \mu + b_u[i] + b_i[j] + P[i] \cdot Q[j]$$

- $P, Q \in \mathbb{R}^{n \times A}$: latent archetypes
- A: number of archetypes

Optimization

Objective function:

$$\sum_{i,j}(R_{ij} - \hat{R}_{ij})^2 + \lambda(\|P\|^2 + \|Q\|^2 + \|b_u\|^2 + \|b_i\|^2)$$

- Optimized using SGD
- Grid search over A and λ

Results: Model Performance

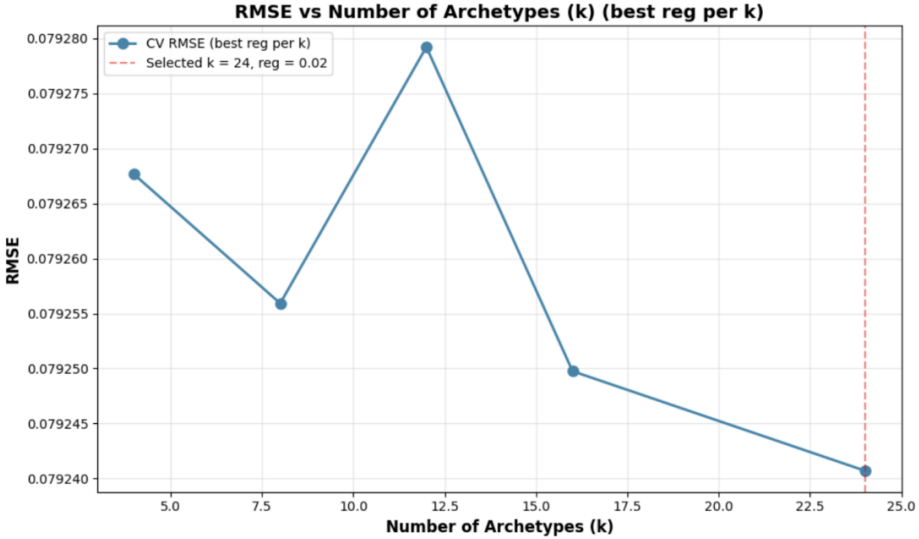
Best configuration:

- Archetypes $A = 24$
- Regularization $\lambda = 0.02$

RMSE:

- Train: 0.0789
- Validation: 0.0796
- Test: 0.0790
- Small validation–test gap $\rightarrow$ good generalization

RMSE vs Number of Archetypes



Latent Archetype Interpretation

- 24 interpretable founder archetypes learned
- Capture patterns across industry, skills, behavior, and roles
- Examples:
  - Risk-tolerant Climate/AI founders
  - Product-focused CPO profiles
  - Technical CTO-heavy archetypes

See heatmaps in appendix for details

Limitations

- No ground-truth outcome data (no success labels)
- RMSE does not directly measure recommendation quality
- Pairwise matching only

Motivation for clustering:

- Team formation beyond pairs (4–5 founders)

Clustering for Team Formation

- PCA used for dimensionality reduction
- Complementarity-based distance metrics
- K-Means clustering ($k = 6$)

Goal:

- Form balanced, multi-founder teams
- Maximize role and skill diversity

Conclusion

- Collaborative filtering provides scalable, interpretable pairwise matching
- Latent archetypes uncover meaningful founder structure
- Clustering extends approach to team-level formation

Future Work:

- Outcome-based validation
- Human-in-the-loop feedback
- Multi-objective optimization