**Project #3: Investor feature engineering and scoring**

***Abstract***

The venture capital landscape is often influenced by prominent early-stage investors who are seen as predictors of startup success. However, focusing solely on well-known names like Elad Gil or Andreessen Horowitz (a16z) limits our predictive capabilities due to data sparsity and overlooks the nuanced features that contribute to an investor's success. This project aims to abstract investor names by creating "investor personas" based on quantifiable features. By doing so, we can enhance our ability to predict startup success and understand how different investor types contribute to it.

***Goals***

1. **Primary goal**: Define investor personas using quantitative features extracted from investment data, thereby reducing data sparsity and improving prediction models for startup success.
2. **Secondary goal**: Analyze the impact of co-investments between different investor personas on the probability of startup success.

***Datasets***

The complete dataset is [here](https://docs.google.com/spreadsheets/d/1m4ngkEJQL8Ysg6KNvrVKZZwWrAuFV4hrCC_nP9jgH0g/edit?gid=0#gid=0).

***Tips***

* investor\_uuid will be the unique identifier to connect data sheets together when you use a dataframe.
* Startup success takes more than 8 years since its founding year. For that reason, we defined quantitative methods to measure success signals.
  + **Outlier success:** Startups founded between 2013 and 2019 that have raised, been acquired for, or IPO'd at $250 million or more.
  + **Interim success:** If a startup that started after 2019 and raised more than $100M, that can be categorized as interim\_success.
  + **Recent success:** If a startup that started after 2022 and raised more than $25M, that can be categorized as recent\_success.

The detailed description of the dataset is below:

* **Co-Investor relationships:** Contains 37,396 U.S.-based seed-stage co-investor relationships. Information on which investors have co-invested in seed-stage startups, along with categories and a quantitative method for identifying outlier companies with three different metrics. Frequency and recency co-investments with others as well as categories should deliver in-depth insights for feature engineering. Take this dataset as the main one to interact with.
  + **Ultimate\_outlier\_success\_250\_mil\_raise (long-term success):** This is the ultimate long term success metric, but this only appears when a startup is at a certain age.
  + **Interim\_success\_100\_mil\_founded\_year\_2019\_or\_above (interim success):** Raising $100M if a company started after 2019 is not an ultimate success, but an interim success that it’s on the track to be an outlier success.
  + **Recent\_success\_25\_mil\_raise\_founded\_year\_2022\_or\_above (graduation rate):** We need to understand the recent success of investors by looking into their startups founded after 2022 to see if older investors remain on track and not to miss new investors. $25M is a solid capital raise for a seed stage startup, which is typically called “graduation” for a seed stage startup to the next financing stage.
* **Long-Term performance:** All U.S.-based seed investments between 2013 and 2022. Includes investors with 20 or more seed investments, totaling 1,011 unique investors. This provides a long-term performance metric of investors by focusing on startups that raised more than $250M, acquired or IPO’ed with a valuation more than $250M.
* **Recent performance:** All U.S.-based seed-stage startups founded after 2019 that have raised more than $25 million. Includes 2,272 unique investors who entered at the seed stage. Offers insights into the recent performance of seed investors in the current market landscape.

***Methodology thoughts***

* **Quantitative methods:** Use quantitative metrics to build features for each investor. Some thoughts are below – you should extend more:
  + First investment year: This would give an idea about how long the investor has been around. This would help us understand if the investor is new or experienced.
  + Annualized number of seed investments: This would give insights on how frequent they invest. This would help us understand if this investor is deploying a “spray and pray” strategy vs “high conviction” strategy.
  + Combine different datasets to calculate outlier success rate , interim success rates such as recent performances.
  + Co-investor success rate: Calculating the average and median success rates of co-investors of an investor could be insightful to understand the network strength of the investor.
  + Focused: Understanding categorical distribution of seed investments, we can understand if an investor is specialized or not.
* **LLM-powered methods:** Using LLMs, company categories can further be defined using category\_list and category\_groups\_list columns. For instance, we can create personas such as “Biotech VC” or “Enterprise-focused VC”.
* **LLM-powered knowledge graphs:** Knowledge graphs can represent investors, startups, industries, and investment outcomes as interconnected nodes and edges, capturing relationships like co-investments, industry focus, and success rates. By analyzing patterns in the knowledge graph, you can identify clusters of investors with similar characteristics, leading to more accurate and nuanced investor personas. You can also use graph-based clustering algorithms (e.g., community detection) to identify groups of investors with similar attributes and relationships.

***Expected Outputs***

**(Primary objective): Calculate the random probabilities**

Understand what random picking probability is. Calculate the average outlier rate, interim outlier rate and recent graduation rate of the whole dataset. In this way, we’ll understand which personas outperform or underperform the index.

**(Primary objective): Investor Personas**

Determine what personas you’ll develop. Develop investor personas such as **L1 Specialist**, **L3 Angel**, or **L5 Generalist**, and calculate their corresponding success rates against the index.

Each persona should be characterized based on features derived from the data. If you want to keep it simple, you can initially focus on developing five different levels: L1 - L5. If you want to add more sophistication, you can add specialization/generalization of the investor via categories AND if the investor is a person or VC firm (person vs company) AND many more personas like this.

*Example Output*

|  |  |  |  |
| --- | --- | --- | --- |
| **Persona** | **Outlier Rate** | **Recent Graduation Rate** | **Number of investors** |
| L3 | 5% | 38% | 30 |
| L2 | 2% | 20% | 20 |
| L1 | 1.5% | 15% | 10 |

**(Secondary objective): Multi-Investor Persona Analysis**

Analyze combinations of investor personas to determine how co-investments affect the probability of startup success. Initially, focus on two-pair combinations, and increase it to three-pair combinations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Investor** | **Outlier Rate** | **Recent Graduation Rate** | **Number of investors** |
| L5, L3 | 20% | 70% | 10 |
| L1, L2 | 10% | 15% | 40 |