A Machine Learning Driven Analysis of Private Equity Funding in Seed-Stage Healthcare Startups

GENE 225 - Spring 2022
Sharan Ramjee

Introduction

- Motivation:
  - The seed stage is extremely crucial for healthcare startups
    - They face high barriers to entry (patent filing, long development periods, etc.)
  - Private equity plays a pivotal role in financing these expenses
    - No information in literature on the factors driving their investments due to risk of exposing their play-books to competitors

- Our Approach:
  - Novel Machine Learning (ML) driven analysis of these factors
    - **Dataset**: Crunchbase
    - **Model**: Gradient Boosted Decision Trees
    - **Technical Approach**: Shapley Additive exPlanations (SHAP)
  - Two step process for analysis:
    - Train ML model on Crunchbase data
    - Use SHAP to probe ML model and gain insights

Dataset

- Collection:
  - Very recent data (May 2022) from Crunchbase
  - Limited to seed-stage healthcare startups headquartered in the United States
  - Tabular dataset with 1,000 examples
    - **Features**: investors, founders, products, patents, etc.
  - Train-test split of 90-10
    - 10% (100 examples) used for test set

- Pre-processing:
  - Data Type Conversion:
    - **Numeric data**: Convert to floats
    - **String data**: Convert to one-hot floats
    - **List data**: Convert to one-hot floats
  - Missing values:
    - **Zeros**: Replace appropriate features with zeros (Number of active products, etc.)
    - **Means**: Replace appropriate features with averages (Website average visits, etc.)

Model

- Architecture:
  - Regression model because output is continuous variable (total funding amount)
  - Need tree-based ML model for human-interpretable results
  - Use grid-search for hyperparameter-tuning
  - Use Mean Average Error (MAE) to evaluate performance
    - Appropriate since the output is a $ value

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Candidates:
  - **Random Forest**: MAE of 3,223,089.82
  - **Gradient Boosted Decision Tree**: MAE of 3,152,521.54
    - Choose this because lower (better) MAE

Technical Approach

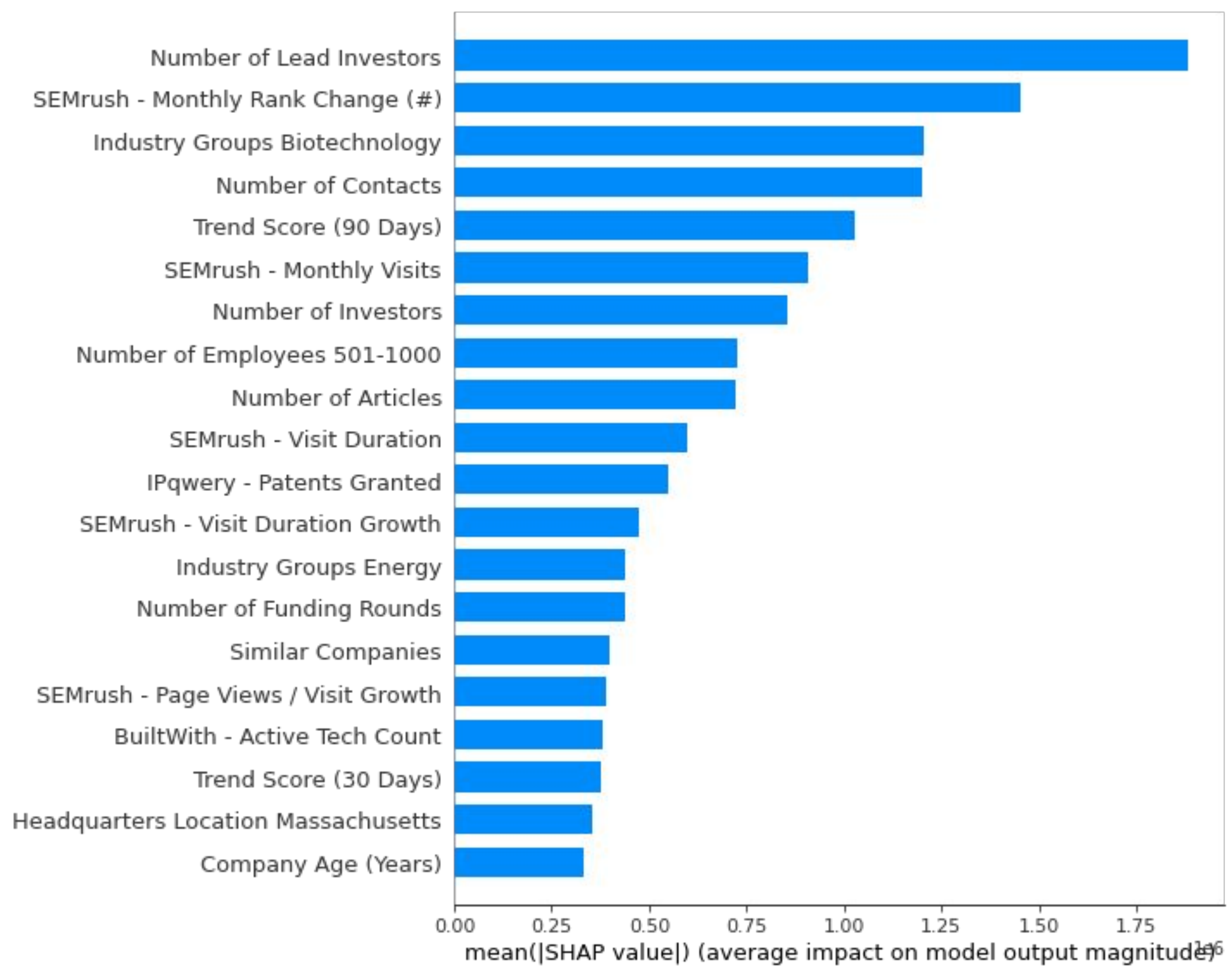
- ML Explainability Method:
  - SHapley Additive exPlanations (SHAP) satisfies several beneficial properties:
    - SHAP scores are a measure of feature importance (rank features)
    - Enforces assumption of independence of features:
      - SHAP scores are a result of causal inference
      - SHAP scores take feature correlations into account

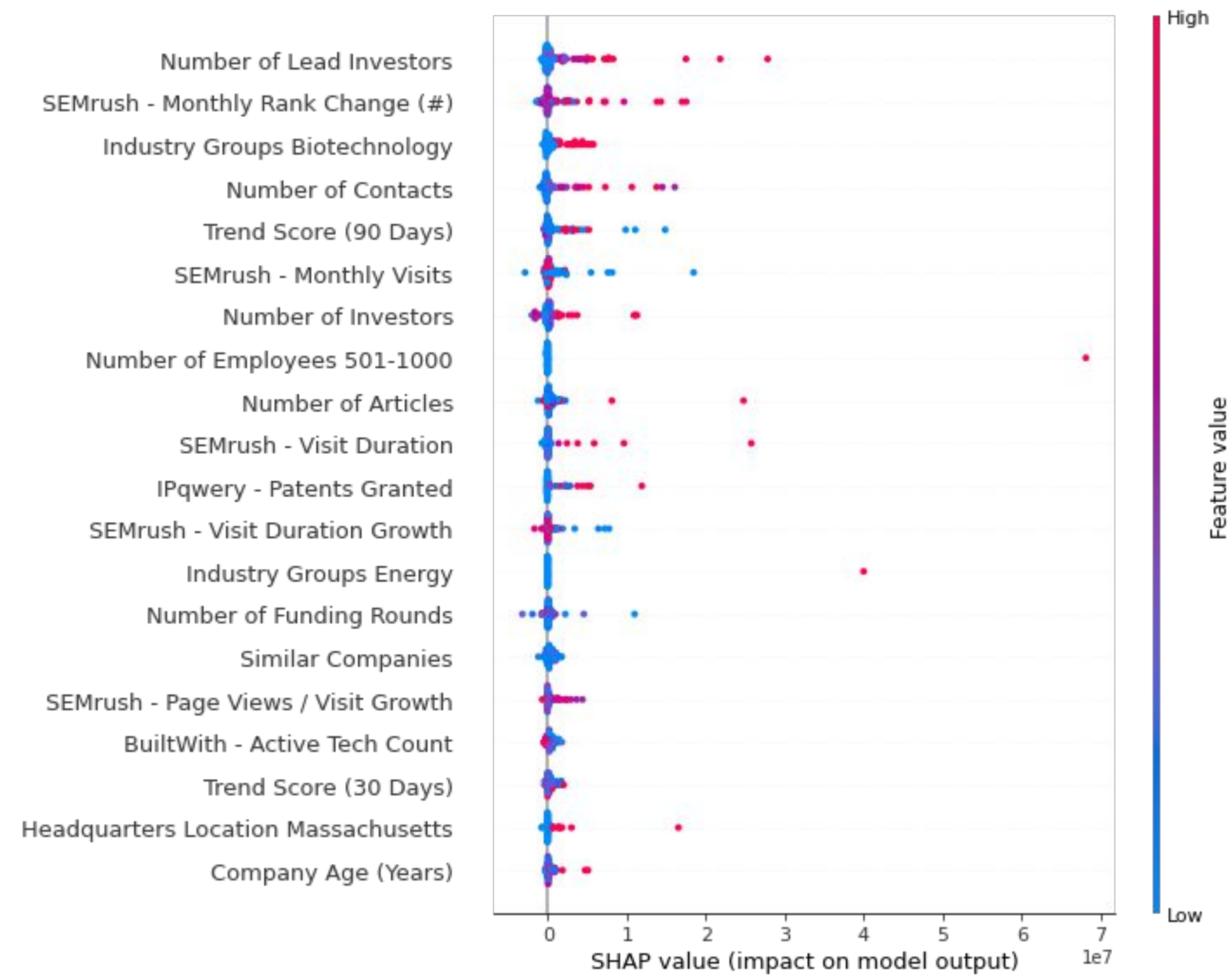
$$SHAP_{feature}(x) = \sum_{set:feature\in set} [|set| \times \frac{F}{|set|}]^{-1}[Predict_{set}(x) - Predict_{set\setminus feature(x)}]$$

- Analysis Types:
  - Some results omitted to keep slides concise
  - **Global**: Over a batch of examples
    - Examine important factors in top and bottom 100 companies with most funding
  - **Local**: Over a single example
    - Examine important factors in top and bottom 3 companies with most funding

Global Analysis

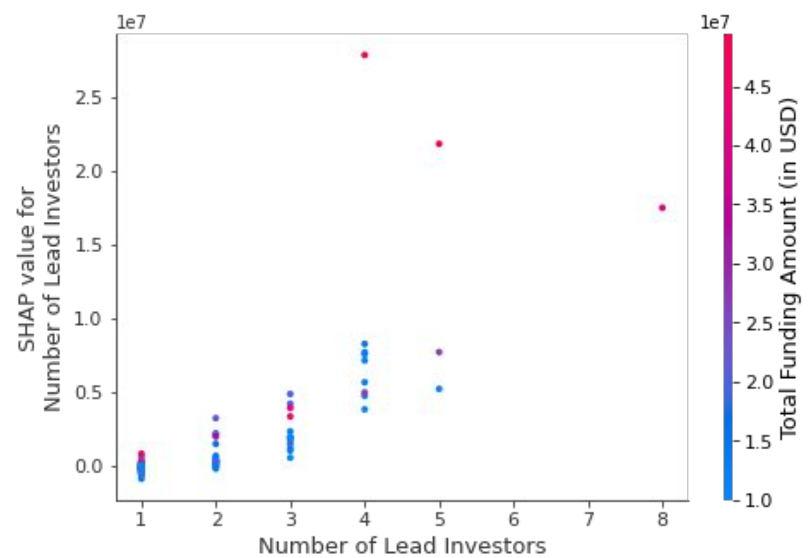
- Top 100 Companies (By total funding amount):
  - 20 most important features by mean(|SHAP value|) on left
  - Individual company SHAP values for 20 most important features on right

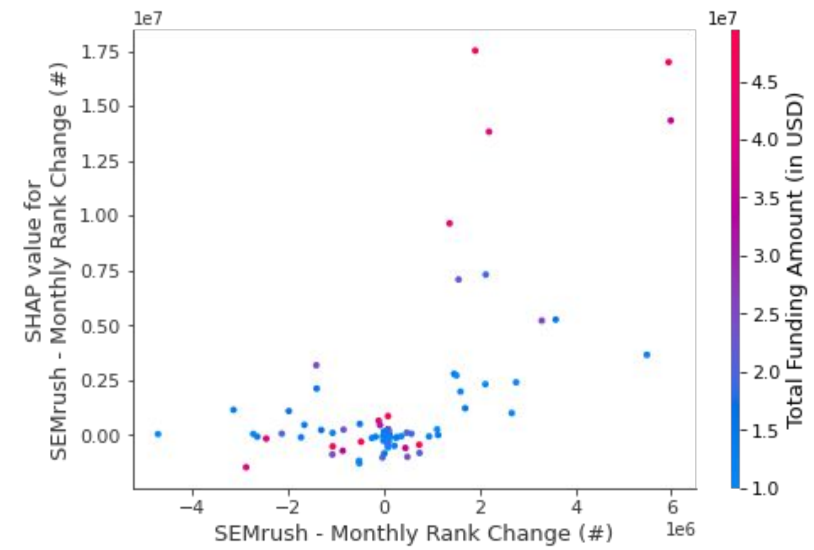


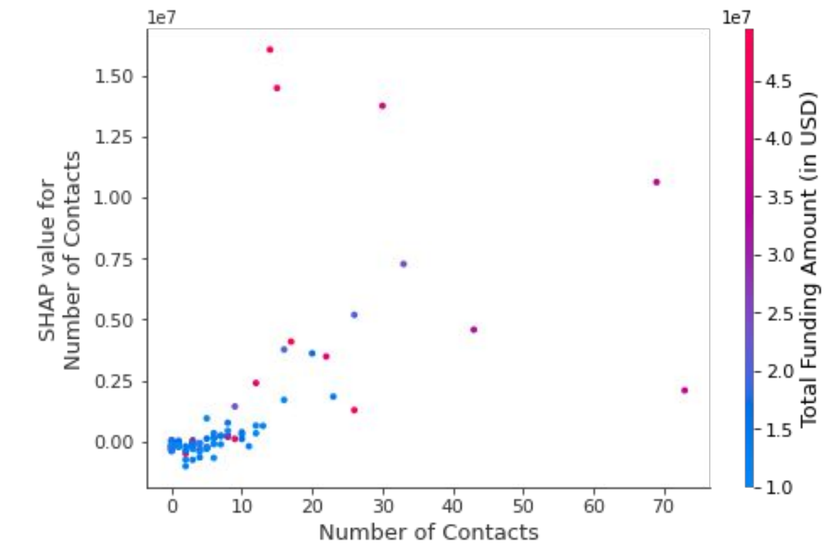


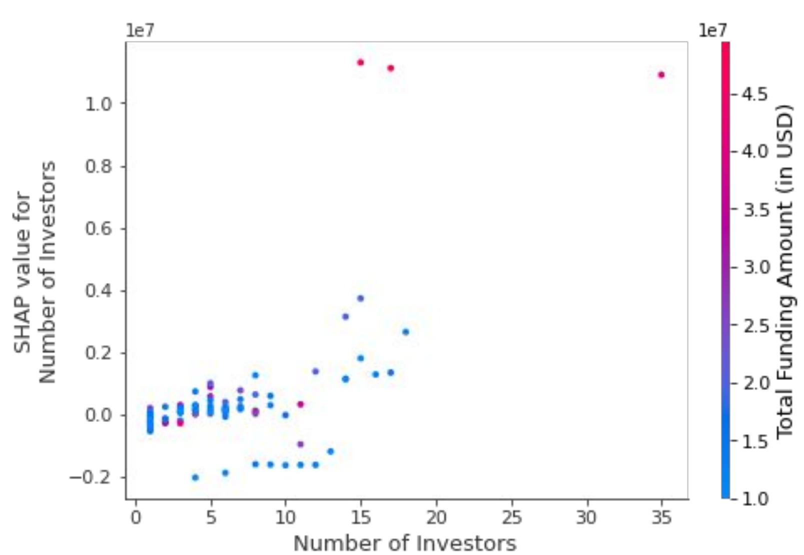
Deeper Global Analysis

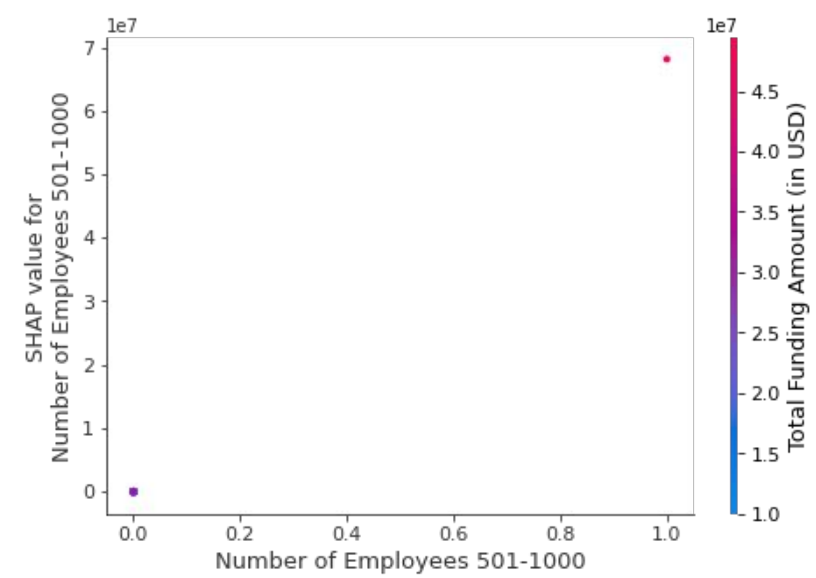
- Dependence Plots:
  - Feature SHAP values vs feature values
    - Color signifies total funding amount raised
  - Plot for most important features
    - Investigate results in literature to support our analysis

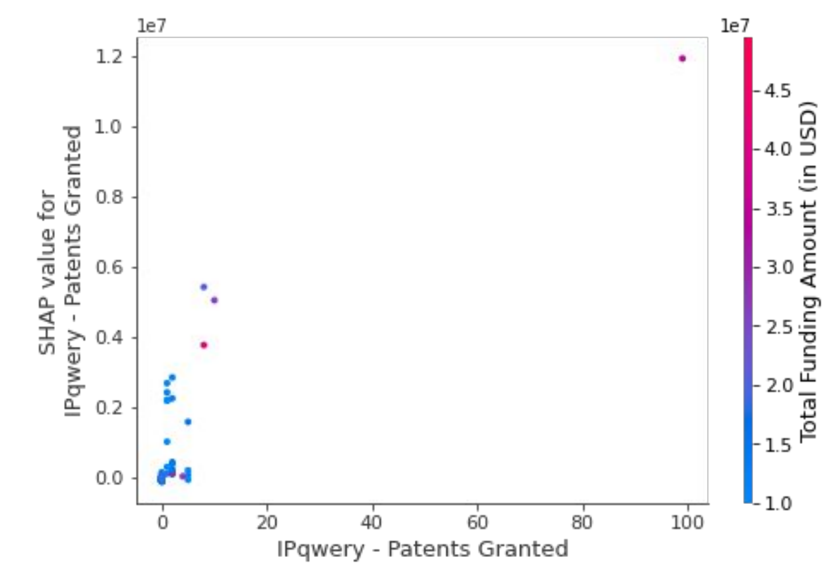












Local Analysis

- Force Plots:
  - ML model prediction of funding raised very close to ground-truth value
    - High fidelity/faithfulness in results
  - **Positive forces** (red and to the right): Make funding amount higher
  - **Negative forces** (red and to the left): Make funding amount lower

- Top 1 Company (By total funding amount):
  - Insightful Science

- Bottom 1 Company (By total funding amount):
  - Gilead Sciences

Summary

- Advantages:
  - **Data**: Used highly recent data for more up-to-date results
  - **Pre-processing**: Used best feature encoding to get most out of features
  - **ML**: Used tree-based ML model that not only achieves impressive performance, but makes results human interpretable
  - **SHAP**: Used SHAP to enforce assumption of independence of features to ensure causal inference (no negative impact due to correlations among features)

- Disadvantages:
  - **Lack of features**: Some critical features that can play a critical role in ML model performance were not available in the dataset. Ex: credibility/track-record of investors
  - **Lack of feature characteristics**: Some features can have both positive or negative influence
    - Ex: Number of articles feature can have positive impact if the articles about the company were positive and a negative impact if these articles were negative

- Overall:
  - Our approach successfully analyzed and evaluated the factors driving private equity investment decisions in seed-stage healthcare startups
    - Results supported by other papers in literature (references in paper)

Thank You!

Code and Results: https://github.com/sharanramjee/healthcare-vc-shap