Funding Forecasts in Biotech:

Insights from Company Characteristics and Economic Indicators

Grace Samuel
(SAMPLE PROJECT)

Abstract

This project explores whether biopharma company characteristics, such as therapeutic area, technology, and development stage, can predict how much funding it will receive. Using company data from DealForma and economic indicators from the Federal Reserve database, five predictive models were built. The best model, XGBoost, was able to estimate funding amounts within a reasonable margin of error. While not exact, the predictions provided directional accuracy, allowing executive search firms to rank clients into high, medium, and low categories for likelihood of funding. Prioritizing companies based on predicted funding enables search firms to make strategic decisions around resource allocation, event attendance, and networking. All work for this project can be found at https://github.com/gsamuel24/DSE6311-repo.git.

Background & Questions

In the context of executive recruitment, insight into the factors that influence biopharma funding is highly valuable. Many biotech companies rely entirely on investor funding for a decade or more as they progress through the drug development process. As Adam (2024) notes, “Unlike other industries, biotech startups often require substantial initial capital to support extensive research and development (R&D), clinical trials, and regulatory approvals before generating any revenue. This high-risk, high reward environment makes navigating the biotech startup funding landscape particularly complex.” Investor funding plays a crucial role in the success of any biotech company, and identifying those with strong financial backing serves as an indicator of their ability to grow and hire new employees.

For executive search firms, understanding where investors are focusing allows for more targeted engagement with stable clients. It also supports long-term relationship building and positions the firm to anticipate and respond to market shifts. If a model reveals trends, such as a rise in gene therapy investment that later shifts to cell therapy, the firm can adapt more quickly than competitors and deliver real-time market insights. Addressing this question has direct implications for business strategy, informing decisions about which companies to prioritize, which conferences to attend, and how to allocate time and resources. This question is considered truly novel, as it reflects a current business challenge being explored within the firm, and no similar models have been identified in previous investigations.

Identified Stakeholders

In this scenario, the Head of Data Science at XYZ Search Partners is the primary analyst, and all stakeholders are internal to the company. The stakeholders include the two partners of the executive search firm, who are non-technical decision-makers and will use the model to drive business strategy, as well as the research team of four people, who are technical and have an advanced understanding of the information. The Head of Data Science will develop the model independently before presenting it to the team. Once developed, the model will be presented to all stakeholders in an information session that includes a high-level summary of the project. Moving forward, the Head of Data Science will meet with the partners and researchers separately due to their differing levels of technical understanding. The model will be updated as new data is collected, with the researchers assisting in long-term maintenance. The partners will

receive monthly updates on the findings initially, with the frequency adjusted based on need and the emergence of new trends.

Hypothesis & Prediction

- Hypothesis: Historical funding data can predict the amount of funding a biotech company will receive by revealing patterns related to therapeutic area, indication, technology type, stage of development, and other features.
- Prediction: Biotech companies developing treatments in high-interest therapeutic areas, indications, technology types, stages of development, and other features will receive higher funding amounts.

Dataset and Initial Cleaning

Most of the data for this project was sourced from DealForma, a biopharma database that provides detailed information on deals and funding activity within the industry. Filters were applied to focus on public biopharma companies categorized by small, medium, large, as well as all private biopharma. Companies in the device, diagnostic, and manufacturing sectors were excluded so that all included organizations are directly working in drug development. Additional filters removed non-U.S. companies, as each country operates under different regulatory frameworks for drug development. In the United States, the Food and Drug Administration (FDA) serves as the governing body. Including data from other countries, each with its own regulatory processes, would limit comparability.

Additionally, the federal funds effective rate and the ten-year treasury yield were exported from the Federal Reserve website and incorporated into the DealForma dataset. The federal funds effective rate reflects the short-term interest rate at the time the deal occurred, and the ten-year treasury yield shows long-term economic sentiment. As stated in Gatlin (2025), “Biotech stocks could become more attractive with interest rates coming down and inflation being curbed.” Integrating these economic indicators with DealForma’s funding data offers macroeconomic context for understanding shifts in investor behavior and funding trends. The citations and final dataset used in this project are listed below:

- DealForma. (2025). Biotech Funding – Custom Export [funding_data]. Retrieved March 22, 2025, from https://www.dealforma.com
- Federal Reserve Board. (2025). Interest rates – Selected historical data. Data Download Program. Retrieved March 30, 2024, from https://www.federalreserve.gov/datadownload/
- Export Link: Google Sheets Export

Initial data preparation involved standardizing column headers to snake_case formatting and dropping columns not used in analysis. Values in the location, stage_at_funding, company_type, and indications columns were reformatted for readability (e.g., “Biopharma – Public Small” was converted to “Public Small”). Data types were reviewed and altered as necessary, with finalized types listed in the details section of Appendix A. Duplicate rows were identified and traced to manual entry errors

in the DealForma dataset, likely due to the same deal being entered twice by different analysts, a known limitation noted by the data provider.

A missingness analysis was conducted and the results are detailed in Appendix B. Variables with low missingness (less than 5%) were retained and imputed after partitioning the data during the pre-processing and feature engineering stage described in later sections. The indications column contained 25.77% missing values. Since it was an important predictor, the missing values were manually collected by searching each company website for the drug development pipeline which includes indication information. The two predictors with 59.21% missingness, lead_investor_this_round and lead_investor_co_type, were removed to avoid introducing bias.

Exploratory Data Analysis (EDA)

A summary table was created to describe the continuous variables in the dataset. Most funding amounts are at the lower end, but a few companies received exceptionally large investments, which skews the averages upward. This pattern is common in biotech, where a small number of companies attract outsized funding. In contrast, economic indicators like interest rates are more evenly distributed and stable. Table 1 provides an overview of distribution shape and central tendency, which will inform subsequent modeling decisions.

	Variable	Mean	Median	Std Dev	Min	Max	Skew	Kurtosis
0	amount	93.375705	30.00	262.077960	0.10	6000.00	9.994265	143.559566
1	federal_fund_effective_rate	1.425270	0.36	1.816393	0.04	5.33	1.193521	-0.049240
2	10yr_treasury_yield	2.463186	2.34	1.047058	0.52	5.25	0.317049	-0.619290
3	total_raised_all_rounds	1388.461033	226.00	6722.688273	1.00	61740.00	7.414527	55.805863

Table 1. Summary statistics for continuous variables, including central tendency, spread, and distribution shape.

A frequency table for stage_at_funding (Figure 1) showed that most deals occur during Phase II (23.3%) and Phase III (17.7%) of development. Early-stage funding (Preclinical/IND and Phase I) is also common, while later-stage or platform-based deals are less frequent. A small share of records fell outside traditional clinical stages.

	Stage at Funding	Count	Percent
0	Phase II	2232	23.3
1	Phase III	1698	17.7
2	Preclinical / IND	1497	15.6
3	Phase I	1420	14.8
4	Approved	1345	14.0
5	Platform / Discovery	1002	10.4
6	Not Disclosed	206	2.1
7	Not Applicable	150	1.6
8	Diagnostic - Any	23	0.2
9	Device - Any	19	0.2

Table 2. Frequency table for the stage_at_funding variable to summarize the distribution of drug development stages at the time of funding.

Next, a correlogram was used to explore relationships between continuous variables. As shown in Figure 1, companies that raised more historically also raised more in the current round, indicating strong collinearity between amount and

total_raised_all_rounds. Similarly, the two macroeconomic indicators are highly correlated. One variable from each pair was dropped to reduce multicollinearity and improve model stability, as seen in the data dictionary in Appendix A.

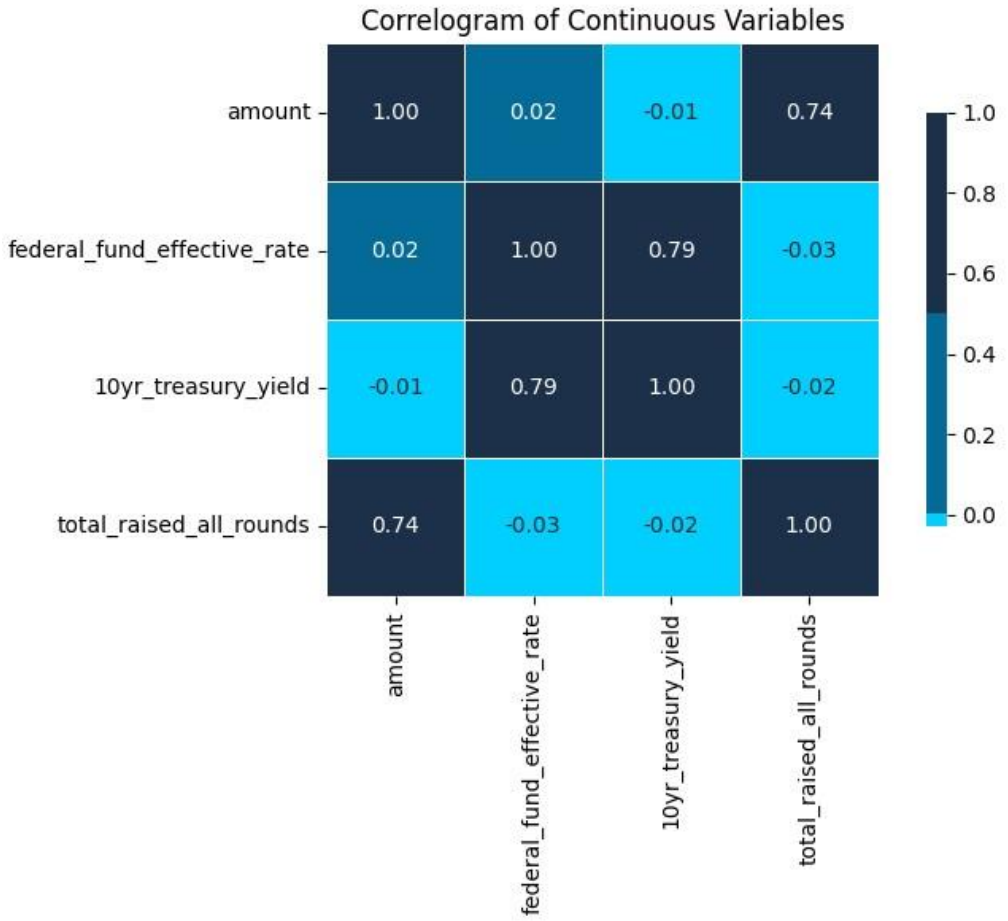

Figure 1. Correlogram of correlations among continuous variables used in the analysis.

The outcome variable was explored to understand the distribution of funding amounts, as shown in Figure 2. To improve visibility of the overall pattern, the top 1% of funding values were removed. Most companies raised under $100 million, while a few received exceptionally large investments, resulting in a highly right-skewed distribution.

To address this, the variable will be log-transformed during the pre-processing and feature engineering phase to stabilize variance.

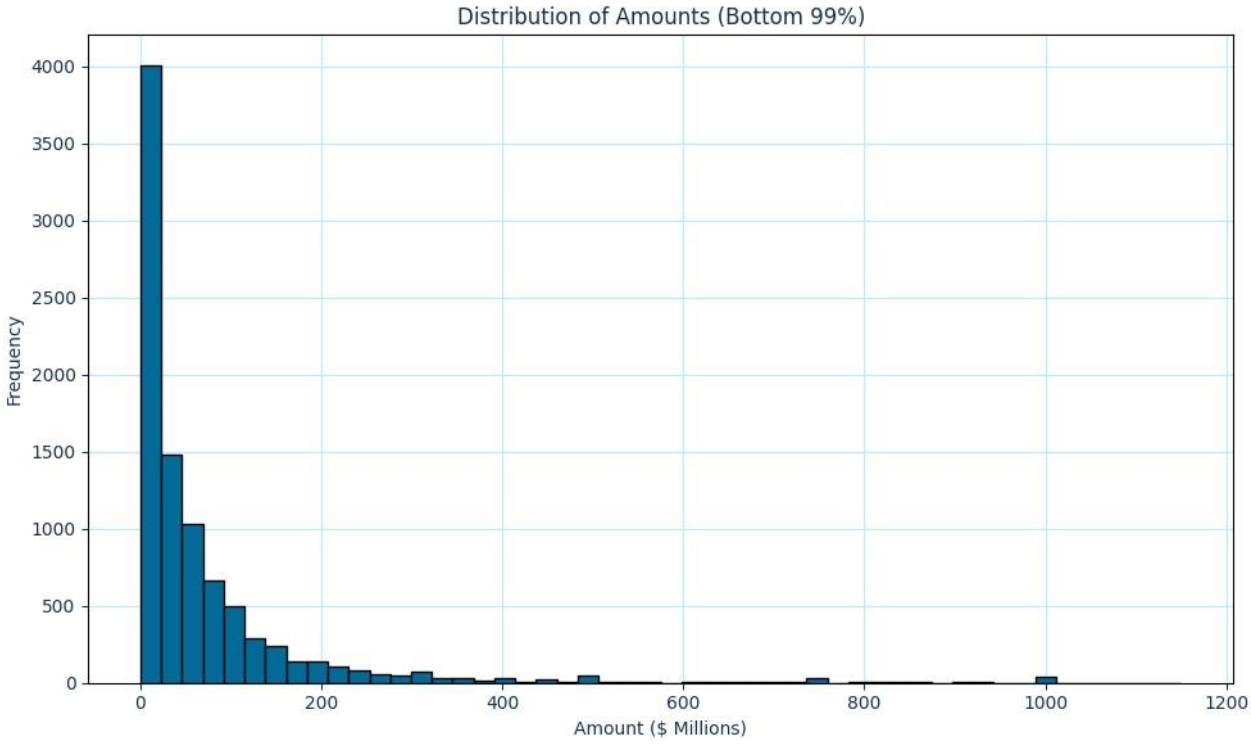

Figure 2. Histogram displaying the distribution of funding amounts (in millions of dollars USD) for the bottom 99% of records in the dataset

The company variable was explored to understand how often each company appears in the dataset, as shown in Figure 3. This bar chart displays the 20 companies with the highest number of records. A small number of companies, such as Amgen, Gilead, and Cel-Sci, appear most frequently, suggesting they are prime targets for investors.

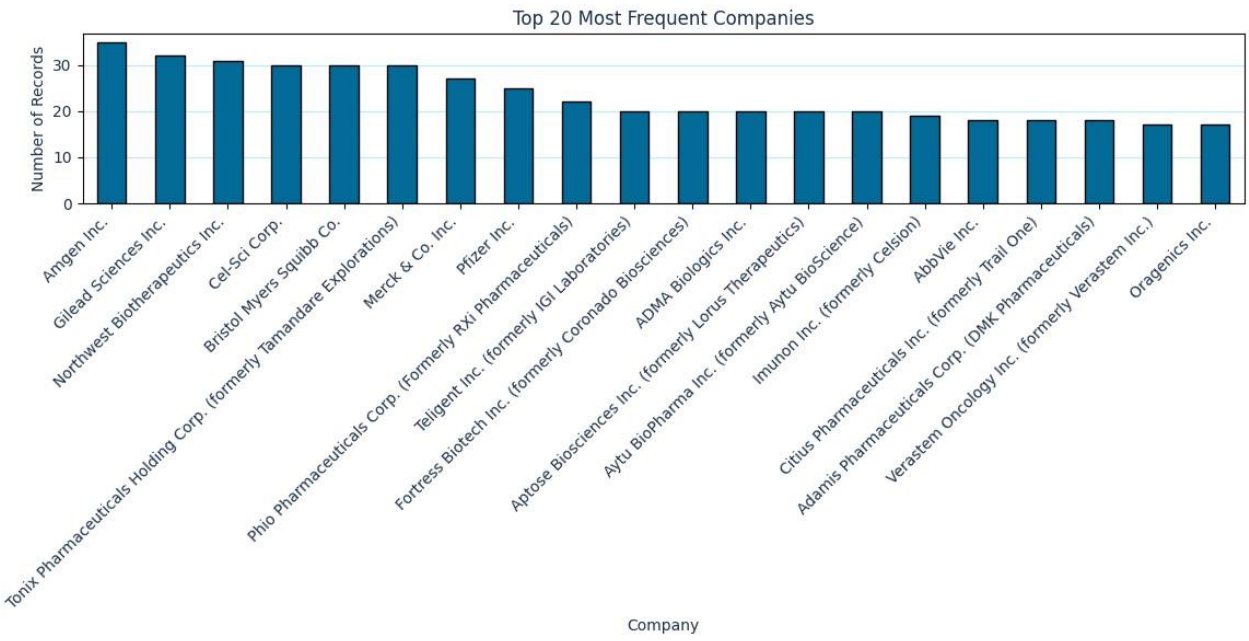

Figure 3. Bar chart displaying the top 20 most frequently occurring companies in the dataset, based on the number of associated records.

Funding trends were analyzed by year and company type to compare public and private deal activity. As shown in Figure 4, public funding was more volatile, peaking between 2014 and 2020 before declining sharply in 2025. Private funding showed steadier growth, with a major increase in 2021.

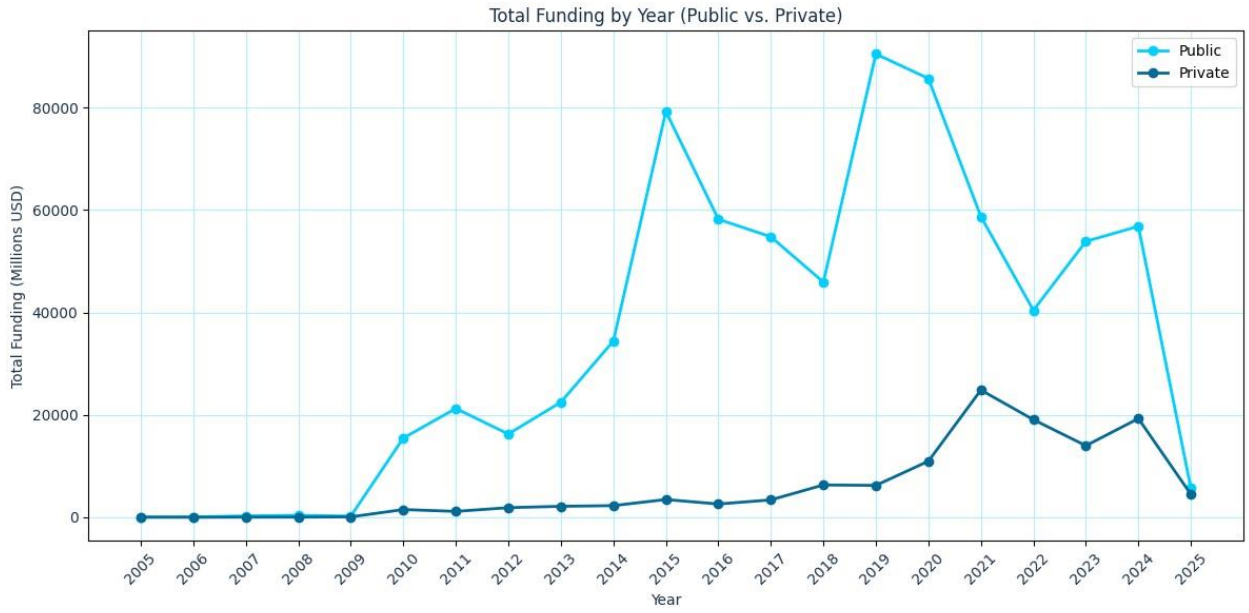

Figure 4. Line graph showing total annual funding for public and private companies. Public funding fluctuated more widely, while private funding followed a steadier upward trend.

Investors were explored to identify those most frequently leading investments in biopharma companies. The most active lead investors were determined by counting how often each appeared in funding rounds. As shown in Figure 5, RA Capital Management LLC led the most rounds, followed by Hercules Capital Inc. and OrbiMed Advisors. The chart highlights a small group of firms that are prominent in the biopharma industry and known for setting funding trends. Future decisions will be made about whether to keep or exclude this variable from further analysis due to missing data in many rows. Options for retaining the data will be explored.

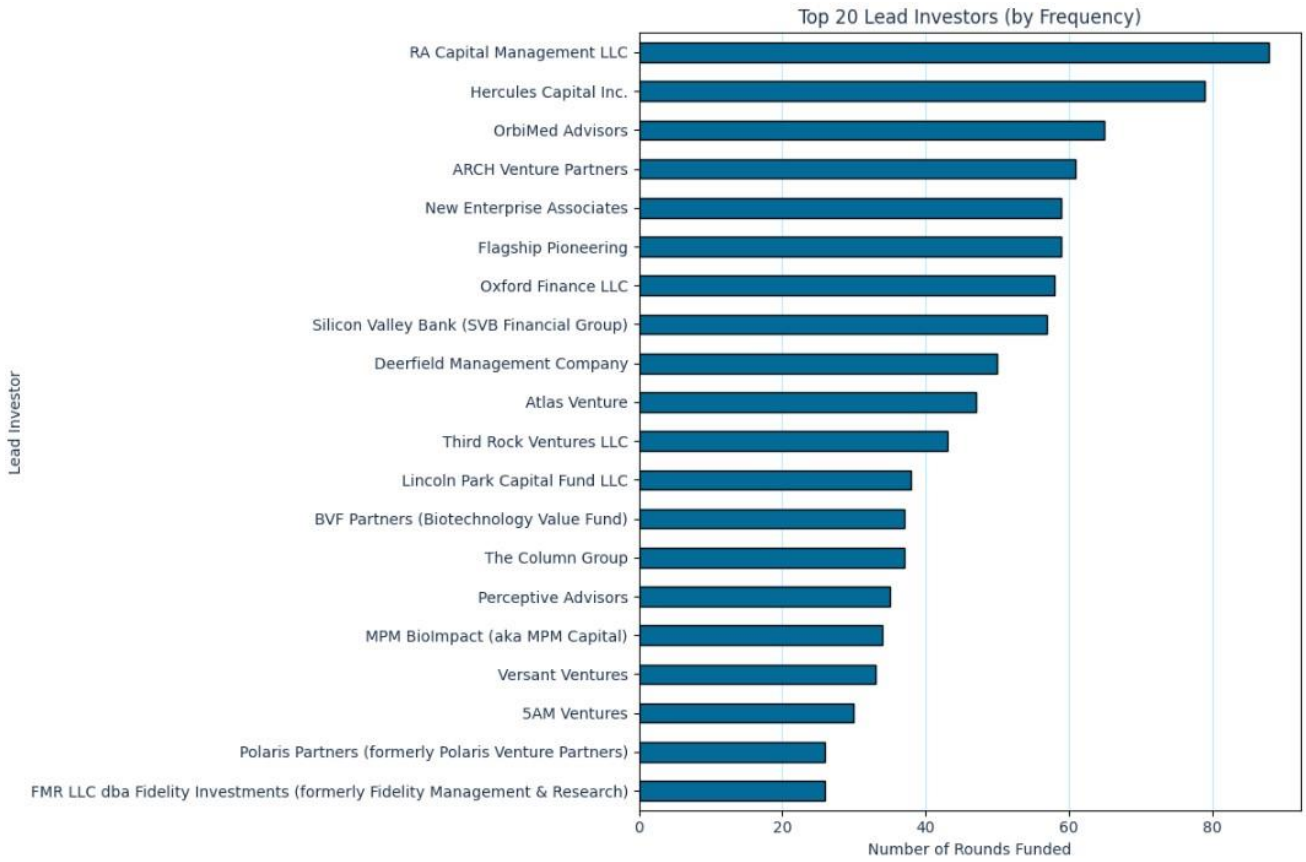

Figure 5. Horizontal bar chart showing the top 20 most frequent lead investors in biopharma funding rounds. RA Capital Management LLC led the most rounds, followed by Hercules Capital Inc. and OrbiMed Advisors.

A bivariate boxplot (Figure 6) of funding amount by company type was created to examine funding patterns. Outliers above the 95th percentile were removed for clarity. Public Large and Medium companies received higher funding on average, while smaller companies showed a higher concentration of lower funding amounts. Because Public Large companies dominate the highest funding values and the stakeholders for this project are focused on small to midsize biopharma firms, Public Large companies will be removed from the dataset during the pre-processing and feature engineering phase.

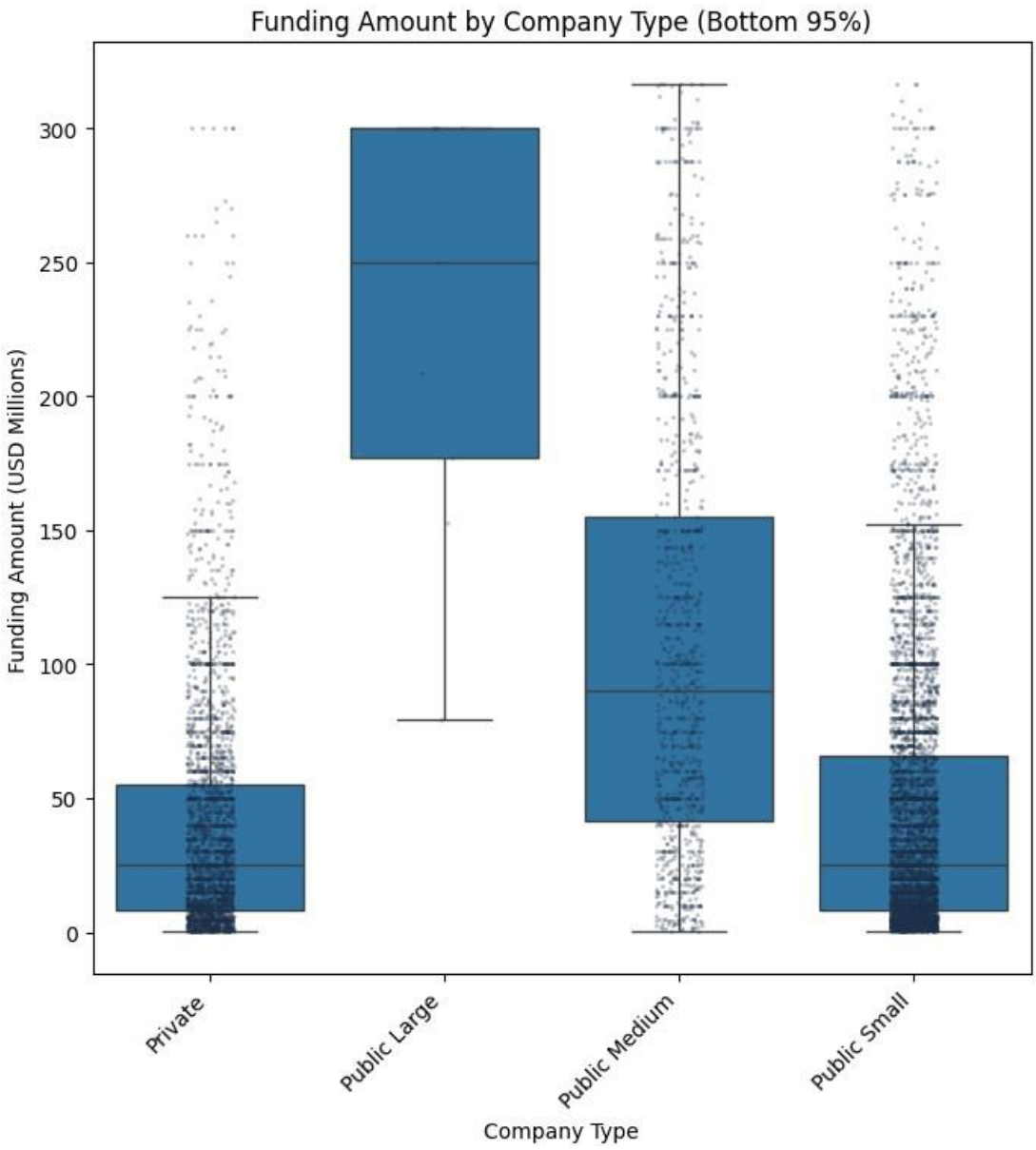

Figure 6. Boxplot of funding amount by company type (bottom 95%), with individual data points overlaid.

Lastly, simple bar charts were made for the remaining predictors to explore the frequency of instances for the top 8, as shown in Appendix C. These charts helped highlight dominant categories within each variable and identify potential class imbalances that may impact modeling.

Pre-Processing and Feature Engineering

Following exploratory data analysis, additional data cleaning was performed to address concerns uncovered through the EDA. Variables with high missingness (lead_investor_this_round, lead_investor_co_type) were dropped due to over 59% missing values, which made reliable imputation unlikely and risked introducing bias. Highly collinear variables (10yr_treasury_yield, total_raised_all_rounds) were also removed to reduce redundancy and improve model stability, particularly for methods sensitive to correlated inputs like regression and PCA. Public large companies were dropped for two reasons: they were few in number and significantly skewed the data, and they are least relevant to the stakeholders, who are primarily focused on small to midsize biopharma firms.

The train/test split was determined using a formula that accounts for the number of predictors in the dataset. The recommended test set size was calculated as $\frac{1}{\sqrt{p}} \times N$, where p represents the number of predictors and N is the total number of records. This method ensures the test set is large enough for reliable model evaluation. The calculation resulted in a test size of approximately 30%, leading to a final 70/30 train/test split. The data was then split into training and test sets before encoding or imputation to prevent data leakage. Encoding or imputing before splitting can allow information from the full dataset to influence the training process, unintentionally leaking patterns from the test set and inflating performance metrics.

Categorical variables were encoded using count encoding, which replaces each category with the frequency of its occurrence in the training set. This method was selected because most categorical variables in the dataset were nominal, had no ordinal relationship, and no two categories had the same frequency. For example, stage_at_funding was reduced to the six most common stages, and all other values were grouped under "Other" before applying count encoding. Similar collapsing and encoding strategies were used for primary_ta, indications, primary_tech, and location. The variable public_private was binary encoded, while round, company_type, and business_model were directly count encoded. See Appendix E for the full Encoding Plan and Appendix D for the resulting Data Dictionary Post EDA.

After encoding, missing data in predictor variables was addressed using a modelbased imputation approach. In the training set, federal_fund_effective_rate and completed_year were the only features with missing values. A Random Forest Regressor was used to predict and fill in missing values for each variable based on patterns learned from the remaining complete features. This method captures non-linear relationships and offers greater accuracy than simple mean or median imputation. The target variable (amount) was not imputed, as doing so could introduce data leakage and compromise model validity. Records with missing target values were removed prior to splitting the data.

Because the target variable (amount) was highly right-skewed, it was log-transformed using log1p to stabilize variance and improve model performance. This transformation was applied after imputation to both the training and test sets. Figure 10

shows a side-by-side comparison of the original and log-transformed distributions of amount. Figure 7 shows that the transformation resulted in a more symmetric, roughly normal distribution. This is helpful for reducing the influence of extreme values, improving the assumptions of regression-based models, and supporting more stable and interpretable model training.

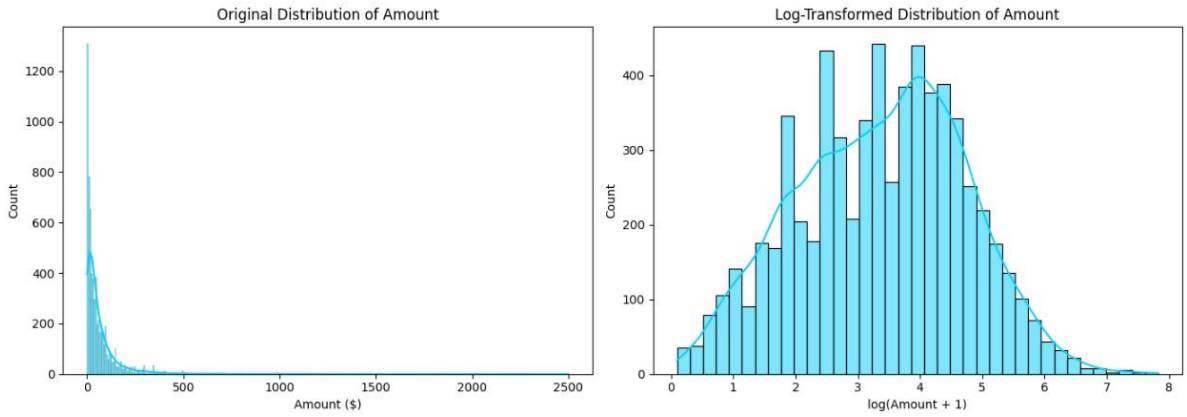

Figure 7. Original and log-transformed distributions of the target variable amount. The transformation reduces right skew and produces a more normalized distribution, supporting improved model performance.

Lastly, all predictor features were normalized using StandardScaler to ensure each variable had a mean of 0 and a standard deviation of 1. This was necessary for downstream techniques like PCA, clustering, and neural network modeling, which are sensitive to feature scale.

Principal Component Analysis (PCA)

The analysis began with unsupervised learning to explore patterns in the data. Principal Component Analysis (PCA) was used to simplify the dataset by reducing the number of features while keeping most of the important information. As shown in Figure

8, the first 10 components captured about 95% of the total variance, meaning the data was successfully compressed without losing much detail.

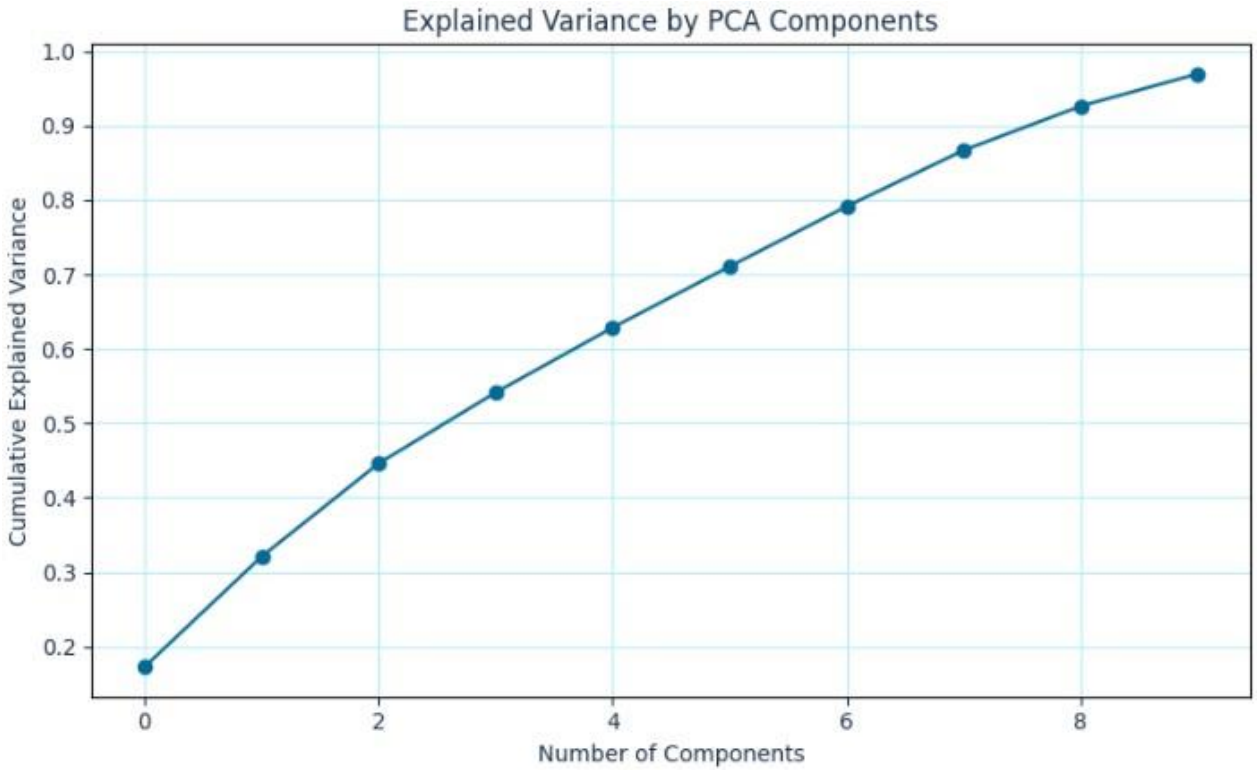

Figure 8. Plot showing the cumulative explained variance from the principal component analysis.

To evaluate the underlying structure in the dataset, K-Means clustering was applied to the PCA-transformed features. An elbow plot (Figure 9) and silhouette scores were used to determine the best number of clusters, with both methods supporting a choice of $k = 4$. However, as shown in Figure 10, the PCA scatterplot revealed a distinct upper grouping of clusters separated by a gap from the remaining data. This separation led to an investigation into the source of the variance.

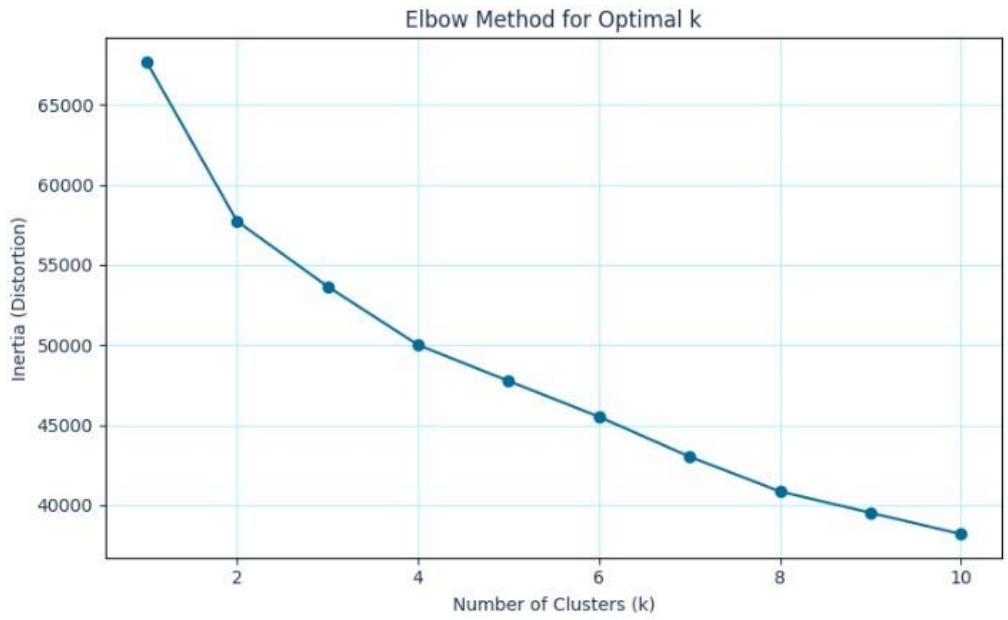

Figure 9. Plot showing the elbow method used to identify the optimal number of clusters for K-Means.

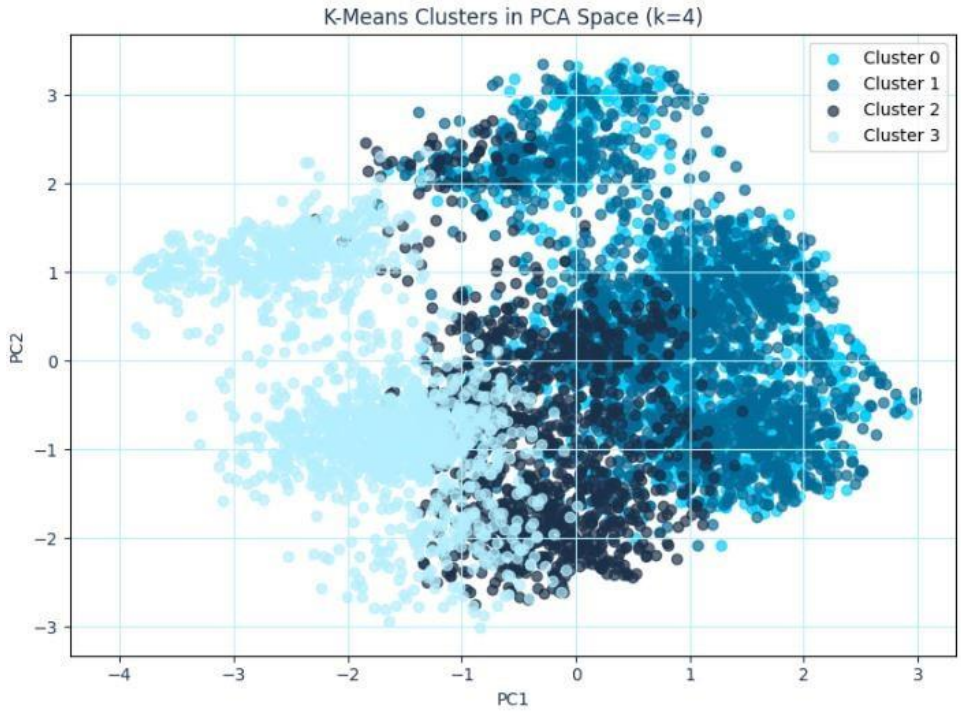

Figure 10. This figure visualizes the K-Means clustering results in two-dimensional PCA space. Each point represents a company, colored by cluster assignment. There is a distinct upper grouping of clusters.

To understand what caused the gap in the PCA plot, the company_type variable was added back to the data and compared to the cluster labels. The results showed that some clusters were almost entirely made up of a single company type, as shown in Table 3. This suggested that company_type was likely the main reason for the distinct separations in the clusters.

company_type	Private	Public Medium	Public Small
cluster			
0	0.008309	0.017804	0.973887
1	0.003079	0.000000	0.996921
2	0.379039	0.620961	0.000000
3	0.967495	0.025494	0.007011

Table 3. This table investigates cluster composition for company_type variable.

This hypothesis was tested by removing company_type from the feature set prior to scaling and PCA transformation. The clustering was then re-run using the same value of $k = 4$. Removing this variable resulted in a more blended and continuous distribution of points across the PCA space, as shown in Figure 11. The previously observed gap was no longer present, confirming that company_type had been the prime influence. This step provided a more balanced view of the remaining predictors' contributions to the clustering structure.

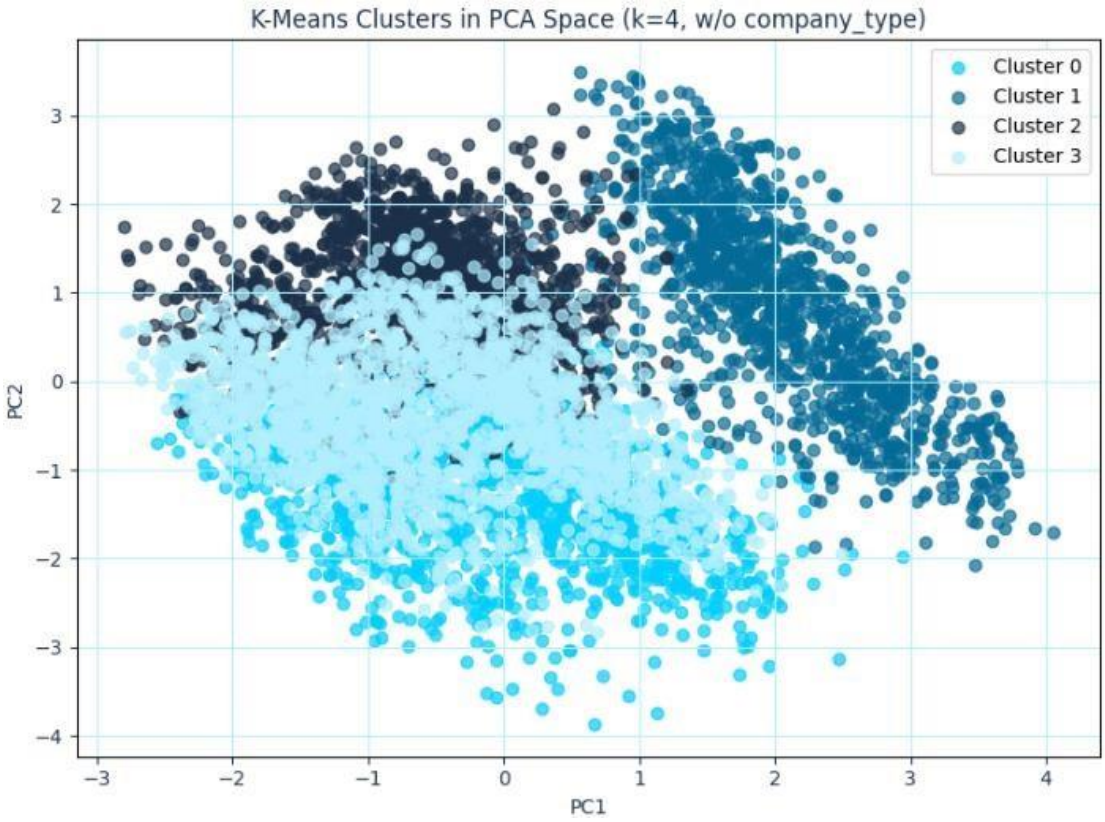

Figure 11. This figure visualizes the K-Means clustering results in two-dimensional PCA space after company_type variable had been removed. The gap observed in Figure 13 is now gone.

Initial Models

Following exploratory data analysis and feature engineering, five supervised learning models were trained to predict company funding amount: Linear Regression, Random Forest, XGBoost, HistGradientBoosting, and a Neural Network (MLP). Linear Regression served as a baseline due to its simplicity, despite restrictive assumptions of linearity and constant variance of residuals. Tree-based models were prioritized for their ability to capture nonlinear feature interactions and their strong performance on

structured data. The MLP model was included to explore predictive accuracy using a neural network architecture.

Each model was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) based on predictions of the log-transformed target. MAE was also converted back to the original dollar scale for interpretation. Cross-validated MAE and RMSE were used to assess model generalization. R^2 was excluded from comparison, as it is sensitive to training set size and model complexity, and less robust than error-based metrics. HistGradientBoosting and Random Forest performed best, both with test MAE around $45 million. XGBoost performed moderately, while Linear Regression and the neural network showed the highest errors, indicating poor fit. Table 4 compares test and cross-validated results for all models. These are baseline results without hyperparameter tuning, which will be addressed next.

Model	**Test MAE**	**Test RMSE**	**Test MAE ($M)**	**Test MSE ($M)**	**CV MAE**	**CV RMSE**
Linear Regression	1.088	1.308	$54.52	$13014.58	1.080	1.300
Random Forest	0.818	1.044	$44.72	$9499.03	0.818	1.051
XGBoost	0.882	1.104	$47.50	$11186.24	0.867	1.086
HistGBR	0.817	1.037	$44.67	$9800.39	0.793	1.009
Neural Network (MLP)	1.048	1.319	$54.46	$11243.29	1.049	1.321

Table 4. Comparison of model performance using test and cross-validated MAE and RMSE. Lower values indicate better accuracy and generalization.

Model assumptions were explicitly addressed. Linear Regression was tested for normality and homoscedasticity using residual plots. The residuals were roughly normally distributed and centered around zero, with mild heteroscedasticity observed but no strong violations, as shown in Figure 12. Tree-based models do not require assumptions of linearity or normality, though they can be sensitive to noise and multicollinearity, which were addressed during preprocessing. The neural network was trained on standardized data, satisfying its assumption of scaled inputs. The dataset meets assumptions of independence and noise reduction, though the sample size may have been too small for the neural network to perform optimally, contributing to its weak results.

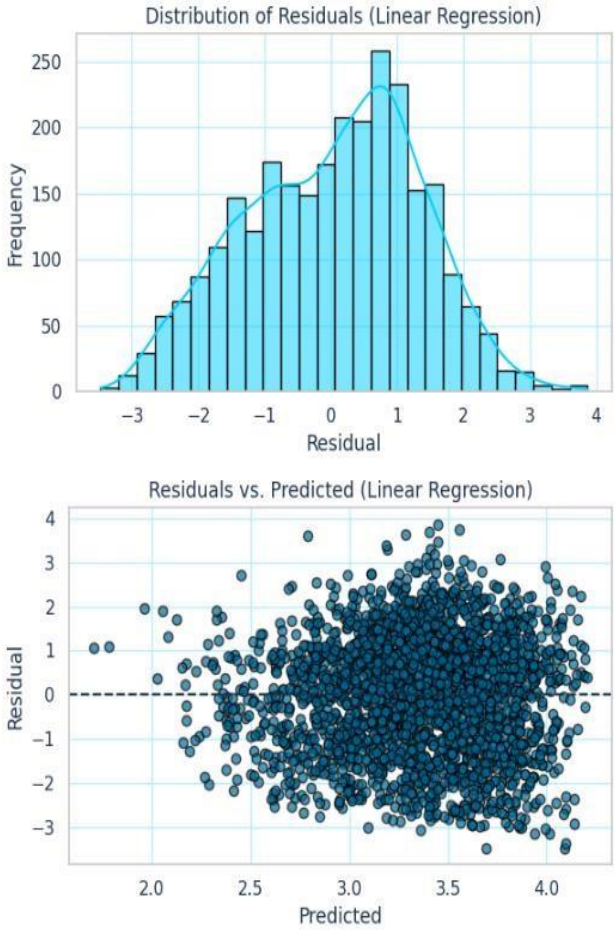

Figure 12. Distribution of residuals and residuals vs. predicted values for the linear regression model. Residuals are approximately normally distributed and centered around zero, with mild heteroscedasticity visible across predicted values.

To control for overfitting, 5-fold cross-validation was performed for all models, evaluating MAE and RMSE on log-transformed targets. HistGradientBoosting achieved the lowest cross-validated errors, followed closely by Random Forest, confirming strong generalization. XGBoost showed moderate generalization, while Linear Regression and the neural network exhibited the highest cross-validated errors, indicating underfitting.

Initial modeling showed that tree-based models, particularly Random Forest and HistGradientBoosting, were the strongest performers, achieving the lowest MAE and RMSE values on both the test and cross-validation sets. These models generalized well and captured nonlinear relationships in the data. In contrast, the neural network and linear regression models underperformed, showing high error rates and signs of underfitting. The original research question, whether company characteristics can predict funding amount, remains appropriate and no major revisions are needed. The company_type variable was removed during preprocessing to avoid dominance effects and improve model balance.

Hyperparameter Tuning

Hyperparameter tuning was performed on all models to improve predictive accuracy and control overfitting. Table 5 summarizes the tuned model performance, comparing test and cross-validated MAE and RMSE for each model. Overall, model

performance improved slightly across tree-based models, while the neural network continued to underperform.

Model	Test MAE	Test RMSE	Test MAE ($M)	Test MSE ($M)	CV MAE	CV RMSE
Random Forest	0.820	1.033	$45.02	$9728.91	0.816	1.033
XGBoost	0.781	1.003	$43.06	$8521.51	0.769	0.988
HistGBR	0.798	1.017	$44.01	$9305.63	0.776	0.994
Neural Network (MLP)	1.024	1.269	$52.54	$12026.98	1.020	1.250

Table 5. Comparison of tuned model performance using test and cross-validated MAE and RMSE. Lower values indicate better accuracy and generalization.

The best performing model after tuning was XGBoost, which achieved the lowest test MAE of $43.06M and a cross-validated MAE of 0.769. This represents an improvement over its initial test MAE of $47.50M and shows good control of overfitting, with cross-validation errors closely aligned to test set errors. HistGradientBoosting and Random Forest also performed well, with test MAE values of $44.01M and $45.02M respectively, and stable cross-validated errors, indicating consistent generalization. These models captured nonlinear relationships in the data effectively and provide reliable estimates of funding amounts within an acceptable margin of error for biopharma funding.

In contrast, the Neural Network (MLP) did not improve meaningfully through tuning. It had the highest test MAE at $52.54M and a cross-validated MAE of 1.020,

suggesting continued underfitting. Given this performance, the neural network is not appropriate for drawing conclusions in this context. It is likely limited by the dataset size and the relatively simple architecture tested.

These results support the hypothesis that company characteristics can predict funding amounts with moderate accuracy, particularly when modeled through tree-based algorithms. While the models are not precise enough for exact funding predictions, they are effective in providing directional estimates, making them suitable for strategic insights into funding potential.

Discussion & Next Steps

This project supported the original hypothesis that biopharma company characteristics can be used to predict funding amounts with reasonable accuracy. After training and tuning multiple models, tree-based approaches, especially XGBoost, delivered the strongest results, achieving a test MAE of approximately $43 million. XGBoost was chosen as the final, top performing model. This models offers directional insight that allows stakeholders at executive search firms to categorize clients into high, medium, or low funding likelihood tiers. The primary limitation is that it cannot predict the exact amount of funding a company will receive.

Although the model doesn’t provide precise dollar-level predictions, its’ value lies in directional accuracy and the ability to compare companies relative to one another. Exact funding amounts are not critical to stakeholders. What matters is the relative level of investment, which is exactly what the XGBoost model helps identify. For example, if

the model shows that cell and gene therapy companies are attracting higher funding compared to those focused on small molecules, stakeholders can shift their business development efforts accordingly by networking with the right people, attending relevant events, and prioritizing outreach in those areas.

The next steps include pushing the XGBoost model to production, ranking all companies the executive search firm engages with, and adding a tier identifier into the CRM so that company value can be quickly assessed by the entire team. Dashboards will be developed to help individual recruiters prioritize their time based on the funding potential of companies that candidates have worked at. The Head of Data Science will meet monthly with the firm's Partners to interpret the model's latest output and guide strategy accordingly. The model will be maintained over the long term and regularly updated as new deal data becomes available from DealForma.

Over time, the goal for this model is to become a core tool in the firm's strategic planning process, offering real-time insights into funding trends and helping the team stay ahead of shifts in the biopharma landscape. By combining internal expertise with data-driven decision-making, the firm will be better positioned to identify high-value opportunities and build long-term relationships with companies shaping the future of biotech.

Code Availability

- https://github.com/gsamuel24/DSE6311-repo.git

References

Adam, J. (2024, July 5). The ABC of biotech startup funding. Labiotech.eu. https://www.labiotech.eu/expert-advice/biotech-startup-funding/

Gatlin, A. (2025, January 13). *Biotech stocks prepare for action in 2025: Weight-loss drugs, AI, and Trump 2.0 are the catalysts*. Investor's Business Daily. https://www.investors.com/news/technology/biotech-stocks-2025-weight-loss-drugs-aitrump/

Appendix A

Data Dictionary – Prior to EDA

Variable Name	Variable Description	Details	Role in Analysis
amount	Amount of funding in USD	Numeric – float64 (Currency)	Outcome Variable
company	Biopharma company name	Category (string label)	Identifier
round	Funding round	Category – One-hot encoded	Predictor
completed	Date the deal was completed	Date – datetime.date	Predictor
federal_fund_effective_rate	Federal fund effective rate at the time of funding	Numeric – float64	Predictor (macroeconomic)
10yr_treasury_yield	Ten-year treasury yield at the time of funding	Numeric – float64 – Removed due to high collinearity with federal_fund_effective_rate	Predictor (macroeconomic)
lead_investor_this_round	Lead investor for that funding instance	Removed from dataset – missingness value too high	Predictor

lead_investor_co_type	Lead investor type (ie., private equity, corp vc)	Removed from dataset – missingness value too high	Predictor
stage_at_funding	Drug development stage at time of funding	Category – One-hot encoded (including 'Unknown')	Predictor
primary_ta	Primary therapeutic area of the company	Category – One-hot encoded	Predictor
indications	Primary indication of the company	Category – One-hot encoded	Predictor
primary_tech	Primary technology of the company	Category – One-hot encoded	Predictor
company_type	Classification of company (ie., – public small)	Category – One-hot encoded (including 'Unknown')	Predictor
location	State the company is headquartered in	Category – One-hot encoded (including 'Unknown')	Predictor
business_model	Business model (ie., early stage R&D only)	Category – One-hot encoded (including 'Unknown')	Predictor
public_private	States if company is public or private	Category – One-hot encoded (including 'Unknown')	Predictor
total_raised_all_rounds	Amount USA raised in all funding rounds	Numeric – float64 (currency) – Removed due to high collinearity with amount	Predictor

APPENDIX B

Missing Data Summary

Variable Name	Missing Count	Missing Percent	Notes
amount	298	3.11%	Small percentage; to be imputed during modeling
completed	70	0.73%	Minimal missingness; retain for now
federal_fund_effective_rate	70	0.73%	Minimal missingness; retain for now
10yr_treasury_yield	330	3.44%	Slightly higher; retain, imputation may be needed later.
total_raised_all_rounds	148	1.54%	Low; retain for now
indications	2472	25.77%	Moderate missingness; manually looked at the pipelines for companies with missing data and filled in the missing data
lead_investor_this_round	5679	59.21%	High missingness; deleted column
lead_investor_co_type	5679	59.21%	High missingness; deleted column
All other variables	0	0.00%	No changes needed

APPENDIX C

Data Dictionary – Post EDA

Variable Name	Variable Description	Details	Role in Analysis
amount	Amount of funding in USD	Numeric – log-transformed target variable (float64); imputed using Random Forest Regressor	Outcome Variable
round	Funding round	Count encoded; nominal variable with 16 unique values	Predictor
completed_year	Year the deal was completed	Numeric (int); extracted from original date field; median imputed	Predictor
federal_fund_effective_rate	Federal fund effective rate at the time of funding	Numeric (float); median imputed	Predictor (macroeconomic)
stage_at_funding	Drug development stage at time of funding	Collapsed to top 6 stages + "Other"; count encoded; nominal	Predictor
primary_ta	Primary therapeutic area of the company	"Other," "Unknown," and "Not Applicable" combined; count encoded; nominal	Predictor
indications	Primary indication of the company	Collapsed to top 10 indications + "Other"; count encoded; nominal	Predictor
primary_tech	Primary technology of the company	Collapsed to top 19 types + "Other"; count encoded; nominal	Predictor
company_type	Classification of company (ie., – public small)	Count encoded; 3 nominal categories	Predictor
location	State the company is headquartered in	Top 5 states retained; others grouped as "Other"; count encoded	Predictor

business_model	Business model (ie., early stage R&D only)	Count encoded; 12 nominal categories	Predictor
public_private	States if company is public or private	Binary encoded (0 = Public, 1 = Private)	Predictor

APPENDIX D

Encoding Plan

Variable Name	Encoding Plan
amount	Numeric target variable – no encoding needed.
round	Count encoded – 16 categories. This variable is nominal, not ordinal, and no two categories share the same frequency.
completed	Numeric – no encoding needed.
federal_fund_effective_rate	Numeric – no encoding needed.
stage_at_funding	Top 6 most common stages were kept; remaining stages were grouped into a 7th category called “Other.” Count encoded. This variable is not ordinal overall due to irregular entries in “Other,” and no two categories have the same count.
primary_ta	“Other,” “Not Applicable,” and “Unknown” were collapsed into a single category called “TA Other.” The remaining 18 categories were count encoded. The variable is nominal, and no two categories have the same count.
indications	Top 10 most frequent indications were retained; the rest were grouped into an 11th category called “Other.” Count encoded. This approach limits dimensionality while preserving signal.
primary_tech	Top 19 most frequent technology types were kept; others grouped into a 20th category labeled “Other.” Count encoded.
company_type	Count encoded – 3 nominal categories with no ordinal relationship and distinct frequencies.
location	California, Massachusetts, New Jersey, New York, and Pennsylvania were retained; all others grouped as “Other.” Count encoded to preserve geographic signal without excessive sparsity.
business_model	Count encoded – 12 nominal categories with no ordinal structure and unique frequencies.
public_private	Binary encode – 2 categories.