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

Sharan Ramjee
sramjee@stanford.edu
Department of Computer Science
Stanford University

Abstract

The seed funding is the first official equity funding stage for startups and arguably the most crucial for those in the healthcare space given the high barriers to entry that they face. Private equity plays a pivotal role in funding healthcare startups, especially in the seed-stage to get them off the ground. However, there is not much information in literature regarding the factors driving their investments due to risk of exposing their play-books to competitors. This paper proposes a novel Machine Learning driven approach to analyzing and evaluating these factors. Our analysis will make use of investment data from Crunchbase pertaining to seed-stage healthcare startups headquartered in the United States to train Gradient Boosted Decision Trees, which are human-interpretable Machine Learning models. SHapley Additive exPlanations, which is a technique for probing such Machine Learning models, will then be applied to the trained Gradient Boosted Decision Trees both at a global scale (top and bottom 100 companies by total funding raised) and at a local scale (top and bottom 3 companies by total funding raised) to reveal insights into the factors that drove private equity investments at these companies. The code is available on GitHub: https://github.com/sharanramjee/healthcare-vc-shap.

1 Introduction

The seed funding [1] is the first official equity funding stage for startups. It represents the first official money that a business enterprise raises to finance its first steps, from market research to product development. When it comes to startups in the healthcare space, the seed funding round is arguably the most crucial [2] due to the high barriers to entry that they face, from filing patents to extensive periods of drug development and testing. Private equity, in particular, Venture Capitalists (VCs) and Angel investors, play a pivotal role in financing these expenses in order to get these healthcare startups off the ground [3]. That being said, there is not much information in literature regarding the factors driving their investments due to risk of exposing their play-books to competitors.

This paper proposes a Machine Learning (ML) driven approach to better understand these private equity investment patterns in seed-stage healthcare startups. ML algorithms are often better than humans at modeling and predicting the outcome of certain tasks. In such cases, ML model explainability methods can be used to probe these models as a means of gaining insights into the factors driving the decisions taken by these models [4]. The goal of this paper is to make use of such ML model explainability methods to perform an analysis of these private equity investment patterns in order to gain insights into the factors that drive these investments in seed-stage healthcare startups. The analysis in this paper will be performed in two steps: (1) Training an ML model (Gradient Boosted Decision Trees) on a dataset consisting of private equity investments, and (2) Using an ML model explainability method (SHapley Additive exPlanations) to understand the decisions taken by the ML model. SHapley Additive exPlanations will be applied to perform both a global analysis and a local analysis where the global analysis will reveal insights on the primary factors driving investment decisions for the top and bottom 100 companies by total funding raised whereas the local analysis will perform a deeper dive into these factors for the top and bottom 3 companies by total funding raised.

2 Related Works

There are several related works available in literature that approach the task of predicting private equity funding in startups using ML. However, each of these papers approach the same task from different perspectives and only a few of them focus on seed-stage healthcare startups headquartered in the United States. Even fewer of these papers make use human-interpretable non-black-box ML models to reveal insights into the predictions made by these models. As for the papers that do, these insights have not been explicitly examined in their respective papers out of fear of revealing their play-books to competitors. Finally, none of the papers make use of ML explainability methods such as SHapley Additive exPlanations to capture, rank, and analyze these investment characteristics in a systematic manner.

The papers by Ünal *et al.*[5], Dellermann *et al.*[6], and Ross *et al.*[7] all make use of an ML model to predict whether or not a startup will successfully exit. While the three papers share similar goals to those of this paper, they differ from our goals in significant ways. All three papers use complex ML models such as neural networks to make predictions as to whether or not a startup will exit (either through an IPO or an acquisition), which is a classification task. The goal of this paper, however, is to use a more human-interpretable tree-based ML model such as Gradient Boosted Decision Trees to reveal insights into the prediction of the total funding received by companies, which is a regression task. Furthermore, the analysis of the former two papers is limited to unicorns, which are startups with a post-money valuation of $1 billion or more whereas the goal of our paper is to focus on seed-stage startups, where private equity investment patters may greatly differ.

The paper by Shrivastava [8] is more aligned with the goals of this paper in the sense that it makes use of a tree-based ML algorithm to predict the investment performance of private equity funding. In particular, the paper uses the Random Forest algorithm to predict the performance of the investments of 500 VCs using a dataset collected from LinkedIn and achieves a test set accuracy of 83.6%. It is important to note that the paper evaluates the post-money performance of private equity investments to observe and evaluate whether these investments were profitable or not, which is a distinct goal from our paper. Furthermore, the paper fails to perform an analysis of the factors that affected the profitability of these private equity investments and is merely a tool to evaluate the performance of investments once they have already been made.

The paper by Wu *et al.*[9] corroborates the approach used in this paper: using ML to reveal insights into private equity investment patters for seed-stage startups. The paper makes use of more than 30,000 data samples collected from several sources such as Crunchbase, Mattermark, and PitchBook to train an ML model to predict whether or not a startup made it to the Series-A round. The paper then explores 400 characteristics of each of the successful deals to finally identify the 20 most important characteristics of seed-stage investments that were most predictive of future success. While the paper uses a similar approach to the one used in this paper at a higher scale (30,000 vs 1,000 deals), it suffers from several drawbacks. Primarily, it fails to examine these important characteristics in the paper, again, out of fear of revealing these insights to competitors. Finally, the analysis is too broad and focuses on all seed-stage startups across the world, where the investment characteristics may greatly differ from those of seed-stage healthcare startups headquartered in the United States.

3 Dataset

The dataset that is used for our analysis is collected from Crunchbase [10], which is a platform that hosts business and investment data on private and public companies. Given that new company data is added to Crunchbase every day, the dataset was collected in May 2022 in order to observe more recent and up-to-date results. Furthermore, given the sheer vastness of data available (1,000,000+ companies), the analysis is limited to seed-stage healthcare companies headquartered in the United States, which leaves us with 5,106 companies. Finally, given the Crunchbase data export constraints, data on only 1,000 companies could be successfully exported. Here, a random train-test set split of 90-10 is applied i.e. 10% (100 instances) of the dataset is used as the test set for the analysis. The dataset contains 1 output feature (total funding amount raised) and 180 input features comprising of all company information accessible on Crunchbase: basic information, schools, industries, team, funding, investors, mergers and acquisitions, events, web traffic, active products, patents, and trademarks. Additional data such as post-money valuation, while available, is not included due to obvious correlation (that will negatively influence the ML model explanations) with the total funding raised during the seed round.

Finally, the dataset is pre-processed to be fed into the ML prediction pipeline. Numeric features are encoded as floating-point values and strings are encoded as one-hot floating-point values. While encoding strings as one-hot values (create separate features for different strings) as opposed to categorical values (create different values under the same feature for different strings) makes the ML explanations less human-interpretable, it enables the ML model to capture patterns pertaining to the same feature in a more uncorrelated manner, thus, allowing for improved performance [11]. This form of feature encoding is especially relevant in the case of Gradient Boosted Decision Trees, where the algorithm takes advantage of uncorrelated features to form the tree splits [12]. Feature values missing for certain instances in the dataset are replaced with either 0s or the means/averages of the remaining values in the corresponding features. The choice of replacing missing values with either 0s or the corresponding means was made depending on what is appropriate for the feature. For instance, it makes more sense to replace the missing values for the `Number of Active Products` feature with 0s whereas it makes more sense to replace the missing values for the `Website Average Visits (6 months)` feature with the means.

4 Model

Supervised ML algorithms can be categorized as either regression or classification algorithms [13]. Regression refers to predicting a continuous output value whereas classification refers to predicting a discrete output value among a set of classes. Given that the output in our task is the prediction of the total funding received by a company (continuous variable), the ML model was structured as a regression model. The option to bucket the output value into discrete classes and structure the ML model as a classification model was available [14]. However, this design choice was not made due to the goal of our ML explainability analysis: we want to gain insights into the factors driving funding amounts up or down, not from one class to another.

There are several evaluation metrics that are suitable for regression tasks [15] : Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). In all cases, a lower metric score (lower error) is better. MSE is typically used in most regression tasks and serves as a better metric for scoring the performance of ML algorithms because it amplifies smaller errors, thus allowing for better error correction when training the ML algorithm [16]. However, it comes with the caveat that the errors are squared. In our case, the output is a $ value and as such, the error would be a ^{2} value, which is not a meaningful quantity [17]. Similar arguments can be made in the case of RMSE [18]. That being said, the MAE was used as the metric for evaluating the performance of our ML model. MAE is defined as:

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

where n is the total number of examples in our dataset (1,000), y_i is the ground-truth total funding raised by company i and $\hat{y}_i$ is the predicted total funding raised by company i.

The ML models were created, trained, and evaluated using XGBoost, an ML library that consists of several highly-scalable ML algorithms [19]. Two types of regression models were considered as suitable candidates for the analysis: Gradient Boosted Tree Regressor model (`xgboost.XGBRegressor`) [20] and the Random Forest Regressor model (`xgboost.XGBRFRegressor`) [21]. Both ML models are tree-based learning algorithms [22]. Tree-based models, unlike neural network models, are not black-boxes. This is because tree-based learning algorithms are essentially cascading `if-else` statements and as such, they allow us to directly probe into the model, enabling a more human-interpretable investigation of the decisions taken by such models [23]. Furthermore, empirically speaking, tree-based learning algorithms usually outperform other ML algorithms on tabular data [24], which is the format of our dataset. The Gradient Boosted Decision Tree model trained on the dataset achieved a test set MAE of 3152521.54 whereas the Random Forest model trained on the dataset achieved a test set MAE of 3223089.82. It is important to note that these were the best performances achieved by both ML models after extensive hyperparameter-tuning performed using a grid search on the model's corresponding hyperparameters [25]. Given that the Gradient Boosted Decision Tree model outperformed the Random Forest model, we chose to use the former for our analysis.

5 Technical Approach

The ML explainability method that was used for the analysis is SHapley Additive exPlanations (SHAP) [26]. SHAP is based on the game theoretically optimal Shapley values, which are the average expected marginal contributions of one feature with respect to the remaining set of features [27]. The SHAP scores outputted by the method are a measure of feature importance, where features with large absolute SHAP scores are more important in comparison to features with small absolute SHAP scores. The global importance (i.e. for the entire dataset and not simply a single example) of features were computed by averaging the absolute SHAP scores per feature across the dataset. The SHAP values are computed as follows:

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

One of the advantages of SHAP that makes it an effective tool to probe ML models is that it uniquely satisfies the property of assumption of independence of features. SHAP enforces this assumption during the computation of feature importance scores to ensure that the ML model predictions made are a result of causal inference, thereby preventing counter-intuitive explanations from arising out of correlations among input features or between any of the input features and the output feature [28]. On the other hand, one of its drawbacks is that it is a reflection of the ML model's ability to predict the outcome. If the ML algorithm is unable to model the relationships among the features in order to accurately predict the outcome, then the feature importances computed using SHAP will also be inaccurate [29]. This is precisely why the dataset was split into a train and test set - the MAE obtained on the test set is a measure of SHAP's fidelity, where a lower MAE represents a higher degree of faithfulness in explaining the model.

6 Analysis

ML explainability methods can be applied at two different levels to analyze the decisions taken by an ML model. The first is the global level, where the feature importances are computed and averaged with regards to the predictions of the model for an entire batch of examples [30]. The second is the local level, where the feature importances are computed with regards to the model's prediction for a single example [31]. That being said, the analysis of the factors driving investment decisions in healthcare seed-stage startups is divided into two sections: a global analysis and a local analysis.

6.1 Global Analysis

The global analysis looks at the factors driving the ML model's predictions from a global perspective in the sense that the important features driving the ML model's prediction of the total funding amount raised by a company is collectively analyzed for the top and bottom 100 companies ranked by total funding amount raised [30]. The top and bottom 100 companies were chosen as two distinct subsets of the dataset to align the analysis better with our goals - we would like to know why investors chose to invest more in certain companies over others.

6.1.1 Top 100 Companies

The 20 most important features that impacted the ML model's prediction outcome is plotted in figure 1. In other words, these were the features with the highest average impact on the model output magnitude i.e. mean(|SHAP value|).

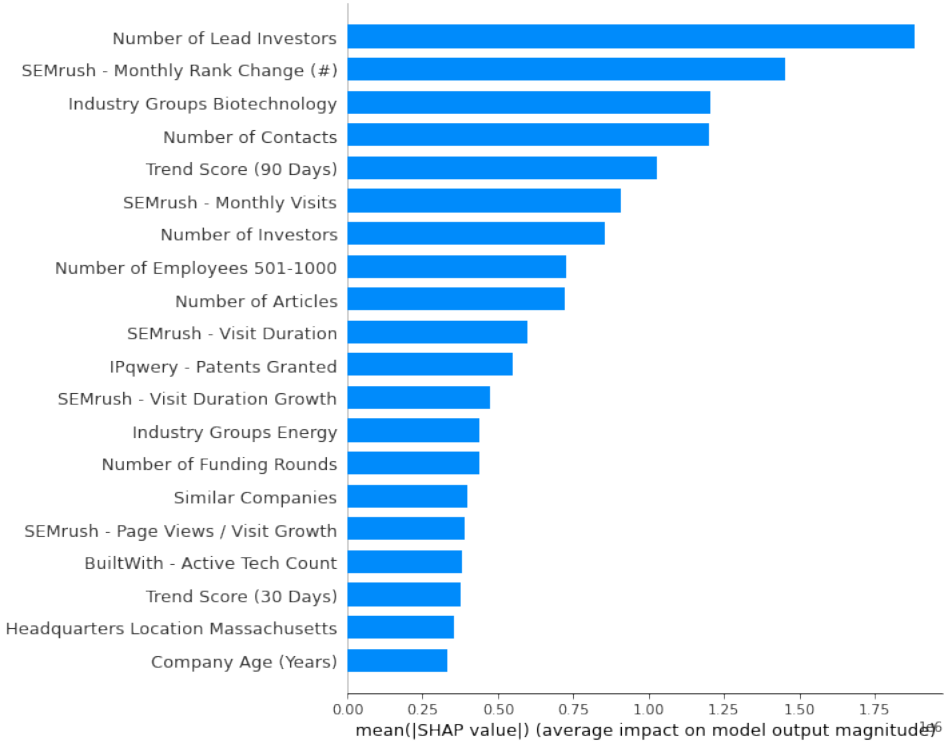

Figure 1: The 20 most important features with the highest mean(|SHAP value|) for the top 100 companies

The SHAP values for these 20 most important features is also computed individually for each of the 100 companies and is plotted in figure 2. In other words, these were the features with the highest impact on the model output magnitude i.e. SHAP value.

Figure 2: The SHAP values for the 20 most important features for the top 100 companies

Number of Lead Investors We observe that as the `Number of Lead Investors` increases, so does its feature importance (SHAP value) and the total funding amount raised by the company. We also observe red points (i.e. companies with high total funding amounts raised) where the `Number of Lead Investors` is low. According to the paper by Li *et al.*[32], the lead investors play a pivotal role in funding, especially in early-stage startups. They found that the greater the number of lead investors, and the greater the credibility of the lead investors, the greater the number of followers in the funding round. Furthermore, they also found that if there are fewer lead investors and they have a higher stake, the fewer the number of followers in the funding round. Given that we lack more specific data on the investors in our dataset, our approach unfortunately does not incorporate the credibility/track-record or stake in the company of the investors involved. That being said, their argument supports our results observed in figure 3 and is essentially what leads to higher funding received by these companies.

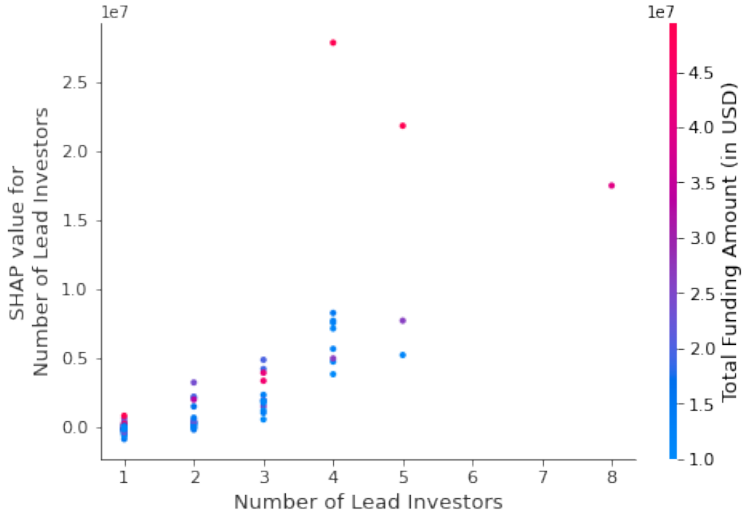

Figure 3: SHAP vs feature values of `Number of Lead Investors` for the top 100 companies

SEMrush - Monthly Rank Change (#) SEMrush is a Software-as-a-Service (Saas) platform that hosts data on keyword search engine searches and online website rankings (including data on search volume, cost per click, etc.) [33]. The `Monthly Rank Change` feature corresponds to how many ranks on average in a month in the context of global website rankings that the company's website rank has changes. Positive values indicate that the website grew in popularity. We observe that as a company's website grows in popularity, so does its feature importance (SHAP value) and total funding amount raised by the company. The paper by Chitkara *et al.*[34] performed an investigation into the importance of web analytics for the success of a startup business. They found that companies that experience explosive growth in website popularity experience more traction from potential investors and the popularity of the website, especially for companies that offer internet/online software-based products, is highly correlated with the success of the companies. Based on our observations, we see that their results substantiate our results observed in figure 4.

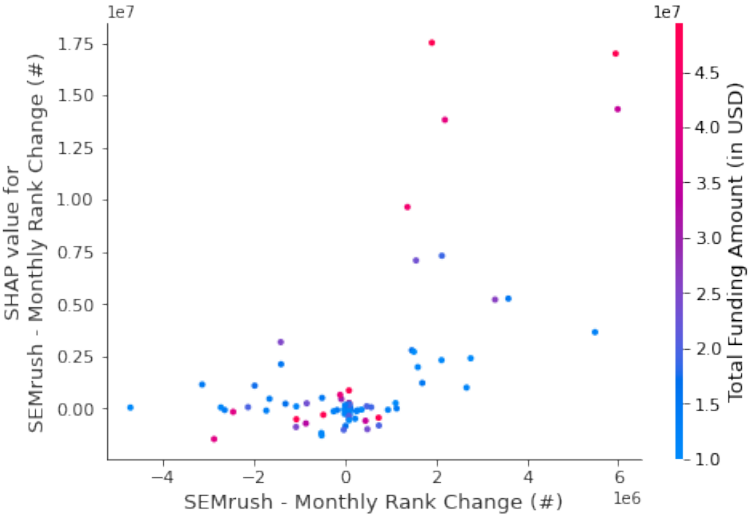

Figure 4: SHAP vs feature values of `SEMrush - Monthly Rank Change (#)` for the top 100 companies

Number of Contacts We observe that as the `Number of Contacts` increases, so does its feature importance (SHAP value) and the total funding amount raised by the company. The `Number of Contacts` feature refers to the number of people in the company that have contact information publicly available on Crunchbase. The paper by Banerji *et al.*[35] performed a study that looks at whether or not more well-connected entrepreneurs are more successful. They found that the "social connectedness" of the founders was the best predictor of the funds raised annually by them for their startups. Furthermore, they found a direct correlation between the number of Crunchbase connections (and number of LinkedIn followers) of the founders and the money their companies had raised, especially in the seed-funding rounds. Again, the results of the analysis in their paper supports the results that we observe here in figure 5.

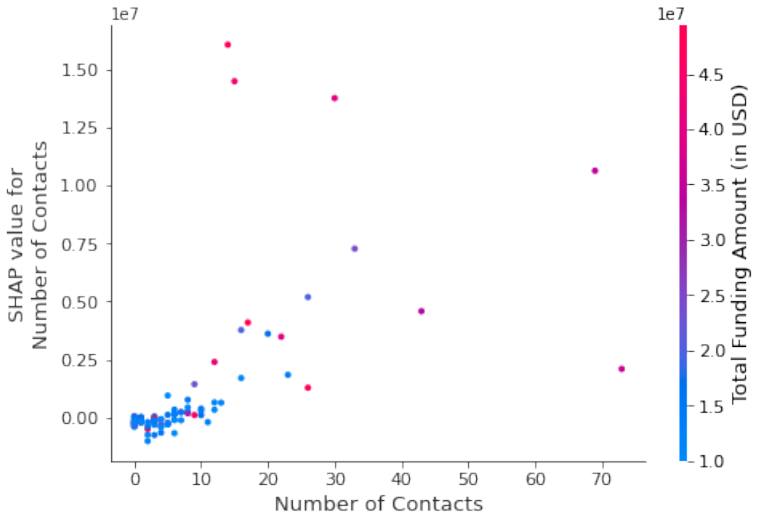

Figure 5: SHAP vs feature values of `Number of Contacts` for the top 100 companies

Number of Investors Similar to the case with the `Number of Lead Investors`, we observe that as the `Number of Investors` increases, so does its feature importance (SHAP value) and the total funding amount raised by the company. Furthermore, we observe a number of red points (i.e. companies with high total funding amounts raised) where the `Number of Investors` is low. We hypothesize that this is the case due to the high investments of extremely wealthy investors or due to the higher stakes borne by the lead investors in the company. The paper by Nanda *et al.*[36] examines the role of various stakes that investments can have on the success of healthcare companies. They found that healthcare companies with a higher number of investors with a higher degree of diversity have a direct impact on the ability of the company to use these funds to overcome the high barriers to entry that these companies face. This analysis supports the results we observe here in figure 6. It is worth noting that this is an obvious result that is derived as a result of the correlation between the `Number of Investors` and the total funding amount raised. However, due to the causal nature of SHAP (as detailed earlier), we can conclude that this is not the case.

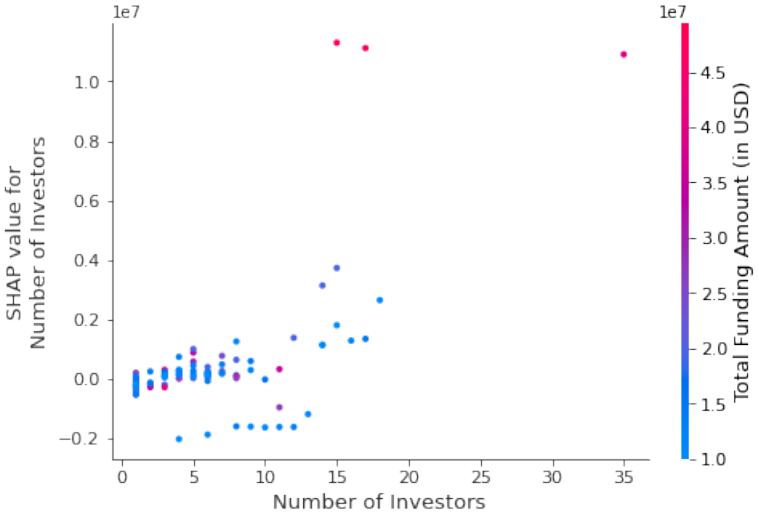

Figure 6: SHAP vs feature values of `Number of Investors` for the top 100 companies

Number of Employees 501-1000 The `Number of Employees` feature was one-hot encoded after forming buckets during the dataset pre-processing step of our analysis. Here, `Number of Employees 501-1000` is the largest bucket and can take on values of either 0 (the company has less than 501 employees) or 1 (the company has 501-1000 employees). We observe that companies that are large enough to have 501-1000 employees have a higher feature importance (SHAP value) and a higher total funding amount raised. The paper by Foster *et al.*[37] evaluates the amount of money early-stage startups receive in the context of their growth and management control systems adoption. They found that companies that have a more systematic management control system in place early on tend to attract more employees, which in turn, allows for them to achieve a "hyper-growth" phase where the company grows rapidly during its early stages. This, combined with a myriad of other factors, tends to bring in more investments from private equity investors and supports the results we observe here in figure 7.

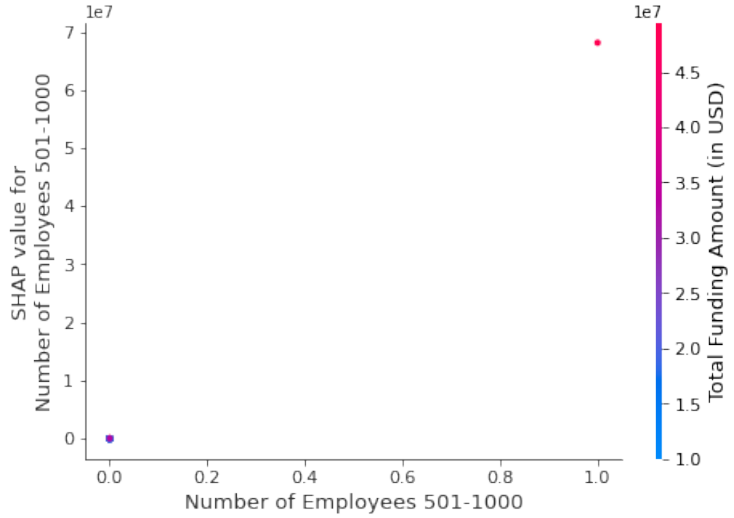

Figure 7: SHAP vs feature values of `Number of Employees 501-1000` for the top 100 companies

IPqwery - Patents Granted IPqwery is a SaaS platform that provides an online software service to query and investigate the intellectual property holdings of companies [38]. The `Patents Granted` feature corresponds to the patents that have been filed, approved, and granted by the United States Patent and Trademark office. We observe that as the number of `Patents Granted` increased, the feature importance (SHAP value) and the total funding amount raised increases. The paper by Conti *et al.*[39] examined the use of patents as signals for startup financing where they construct a model where technology startups use the number of patents they file as a signal for private equity investors. They found that companies with more patents granted tend to attract more traction with such investors and can lead to higher investments in the early-stages of the startup, especially in the field of healthcare. It is worth noting that the reverse is also true: healthcare startups that receive more funding have the money to hire more intellectual property lawyers and file more patents and as such, the number of patents granted is and an important indicator of the success that these companies achieve. Again, their investigation supports the results we observe here in figure 8.

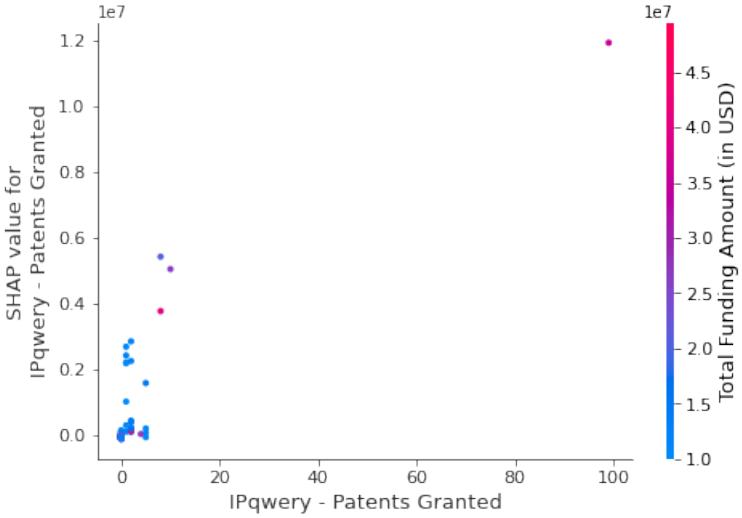

Figure 8: SHAP vs feature values of `IPqwery - Patents Granted` for the top 100 companies

Industry Groups Energy Similar to the case with the `Number of Employees` feature, the `Industry Groups` feature was also one-hot encoded during the dataset pre-processing step of our analysis. Here, the `Industry Groups Energy` feature can take on a value of either 0 (the company does not belong to the energy industry) or 1 (the company belongs to the energy industry). Upon further examination of this feature, we see that these companies primarily work on energy-efficient medical devices such as wearables and implants. We observe that healthcare companies that work in the energy space have a higher feature importance (SHAP value) and a higher total funding amount raised. The paper by Kivisaari *et al.*[40] examines the private equity investment trends in the medical device healthcare sector. They found that companies working on medical devices raised a total of over $80 billion in 2020, which was a sixfold increase from the money raised by such companies in 2019. The medical device space is growing rapidly with new startups emerging and increasing amounts of private equity funding that is being raised and the results of their analysis support the results we observe here in figure 9.

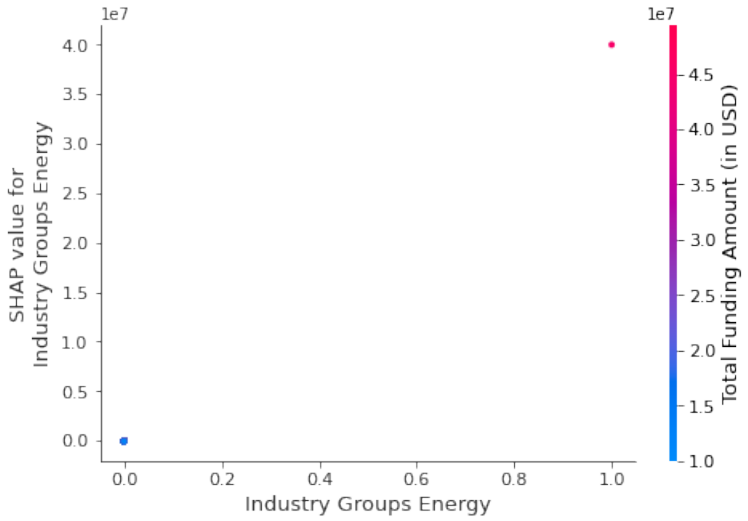

Figure 9: SHAP vs feature values of `Industry Groups Energy` for the top 100 companies

Headquarters Location Massachusetts Again, similar to the `Industry Groups` feature, the `Headquarters Location` feature was also one-hot encoded during the dataset pre-processing step of our analysis. Here, the `Headquarters Location Massachusetts` feature can take on a value of either 0 (company is not headquartered in Massachusetts) or 1 (company is headquartered in Massachusetts). We observe that healthcare companies that are headquartered in Massachusetts have a higher feature importance (SHAP value) and a higher total funding amount raised. The paper by Adler-Milstein *et al.*[41] performs a survey of Boston, Massachusetts based healthcare startups and the private equity funding they have received in the past decade. They found that the extremely high concentration of medical institutions, universities, and VCs based-out of Boston create an environment where healthcare startups can flourish. Several notable healthcare startups that have come out of Boston, Massachusetts (PathAI, Benchling, etc.) attribute their success to this environment and this corroborates the results that we observe here in figure 10.

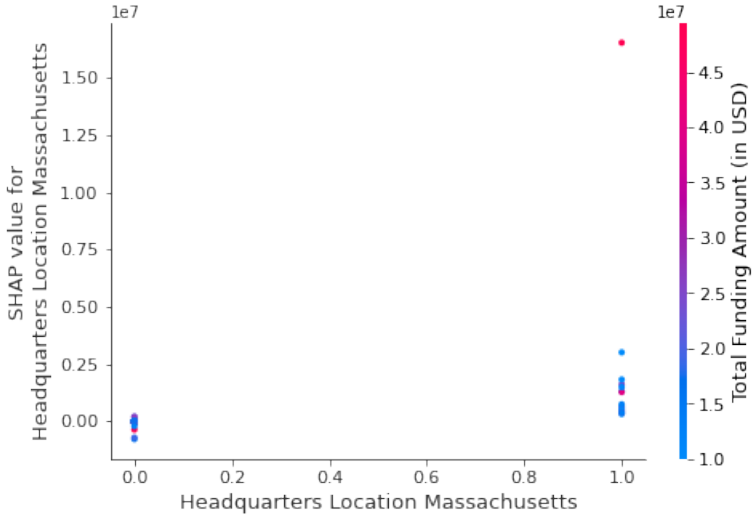

Figure 10: SHAP vs feature values of `Headquarters Location Massachusetts` for the top 100 companies

6.1.2 Bottom 100 Companies

The 20 most important features that impacted the ML model's prediction outcome is plotted in figure 11. In other words, these were the features with the highest average impact on the model output magnitude i.e. mean(|SHAP value|).

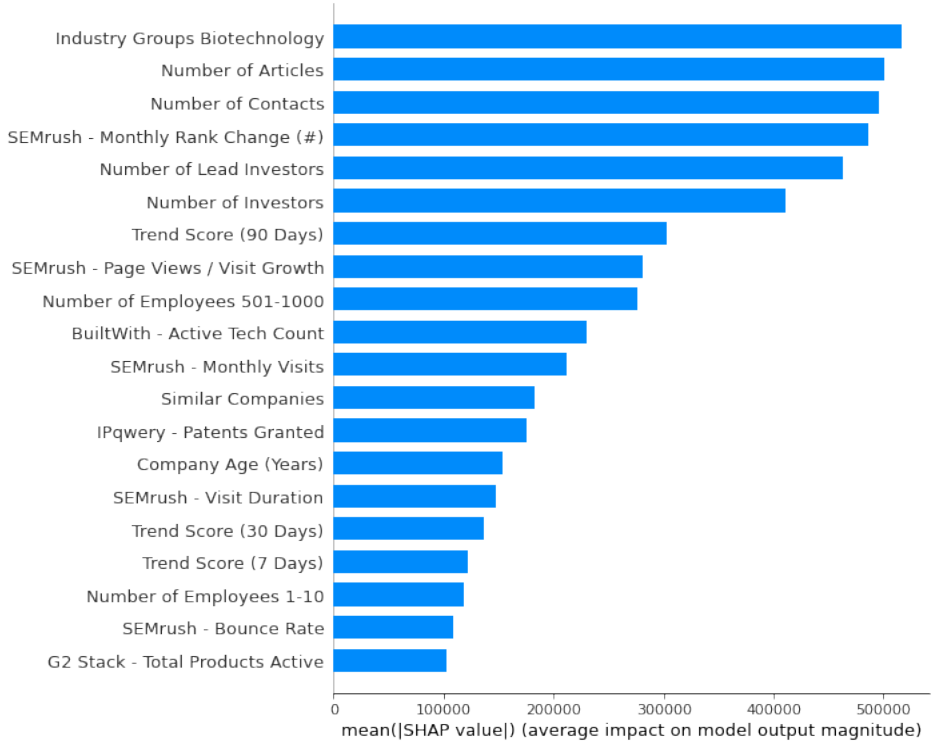

Figure 11: The 20 most important features with the highest mean(|SHAP value|) for the bottom 100 companies

The SHAP values for these 20 most important features is also computed individually for each of the 100 companies and is plotted in figure 12. In other words, these were the features with the highest impact on the model output magnitude i.e. SHAP value.

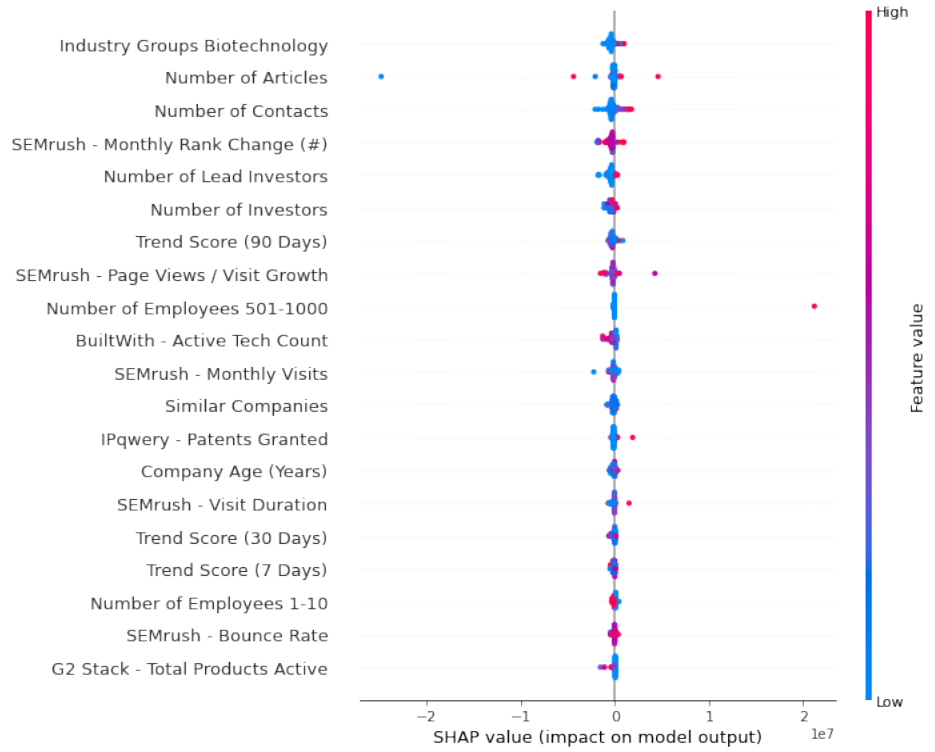

Figure 12: The SHAP values for the 20 most important features for the bottom 100 companies

Given that the goal of our analysis is to gain insights into the factors driving private equity investments in highly successful companies, we have omitted a deeper dive into the factors influencing these decisions for the bottom 100 companies by total funding amount raised. That being said, the 20 most important features contributing to the predictions made by the ML model are highly similar among the top and bottom 100 companies.

6.2 Local Analysis

The local analysis looks at the factors driving the ML model's predictions from a local perspective in the sense that the important features driving the ML model's prediction of the total funding amount raised by a company is individually analyzed for the top and bottom 3 companies ranked by total funding amount raised [31]. As mentioned earlier, the top and bottom 3 companies were chosen as two distinct subsets of the dataset to align the analysis better with our goals - we would like to know why investors chose to invest more in certain companies over others. In each of these cases, we will look at the SHAP force plots to delve deeper into the factors driving private equity funding either higher (red and to the right) or lower (blue and to the left).

6.2.1 Top 5 Companies

Insightful Science Insightful Science is large-sized healthcare startup headquartered in San Diego, California that focuses on providing software to support scientists with research procedures. We observe in figure 13 that the funding amount predicted by our ML model is extremely close to the ground-truth funding amount raised of $100 million and as such, our results have a high degree of fidelity to them. Furthermore, we observe that the two main factors driving the funding raised higher are the `Number of Employees` and the `Number of Articles`. As examined earlier, large-scale companies that have experienced "hyper-growth" tend to attract more investment from prospective investors [37]. Furthermore, the high number of articles is an indication that the company is highly popular as this, again, helps bring in more investors that want to like to turn a profit.

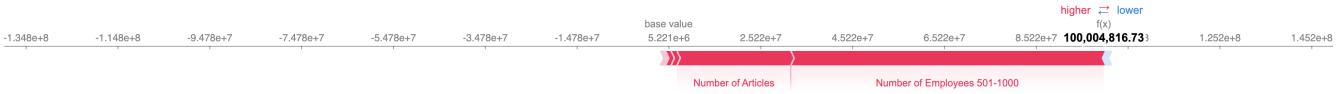

Figure 13: The SHAP force plot for Insightful Science

Avilar Therapeutics Avilar Therapeutics is mid-sized healthcare startup headquartered in Waltham, Massachusetts that focuses on extracellular protein degradation for pharmaceutical purposes. Again, we observe in figure 14 that the funding amount predicted by our ML model is extremely close to the ground-truth funding amount raised of $60 million and as such, our results have a high degree of fidelity to them. In comparison to the previous case, we observe that there are lot more factors driving funding raised higher. Primarily, these are the `Monthly Rank Change` and the `Headquarters Location` features. Both of these features have been examined in detail with regards to the influence they have on the total funding amount raised by companies. A large proportion of Avilar Therapeutics' success can be attributed to the high number of extremely well-educated MD-PhDs that work at the company, most of which come from universities based out of Boston (a simple LinkedIn search reveals that most of these employees came from Harvard University and Massachusetts Institute of Technology). Again, this supports our earlier analysis of the success of healthcare companies that are based in the highly healthcare and biotechnology startup focused environment of Boston, Massachusetts [41].

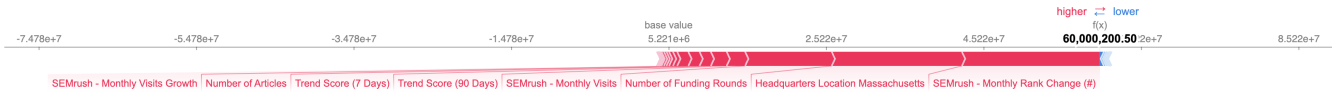

Figure 14: The SHAP force plot for Avilar Therapeutics

Justpoint Justpoint is a small-sized heathcare startup headquartered in New York City, New York that uses AI technology to search for attorneys to help with litigation for health related injuries (car accident, workplace accident, etc.). Again, we observe in figure 15 that the funding amount predicted by our ML model is extremely close to the ground-truth funding amount raised of $58.6 million and as such, our results have a high degree of fidelity to them. We observe that the main features driving funding raised higher are the `Number of Lead Investors`, `Number of Investors`, and the `Global Traffic Rank`. Justpoint has a total of 17 investors with 5 lead investors. As examined earlier, having more lead investors take charge of raising capital attracts a lot of followers, which in turn leads to higher funding raised by companies [32].

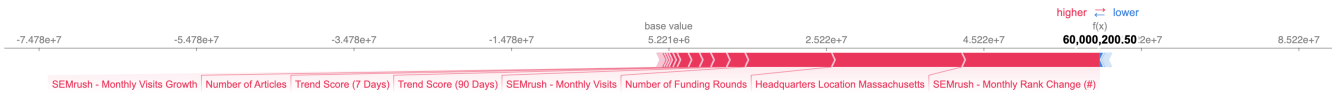

Figure 15: The SHAP force plot for Justpoint

6.2.2 Bottom 5 Companies

Given that most features seen in the case of the top 5 companies drove funding raised higher, we will analyze the bottom 5 companies with a primary focus on analyzing the factors that drive funding raised lower.

Gilead Sciences Gilead Sciences is large-sized healthcare startup headquartered in Foster City, California that provides biopharmaceutical services for discovering, developing, manufacturing, and commercializing therapies for critical diseases. We observe in figure 16 that the primary factors driving funding raised lower are the `Bounce Rate` and the `Monthly Rank Change`. The `Bounce Rate` feature refers to the number of times website access was denied or "bounced" away to another site. A high `Bounce Rate` is negative because it can turn away potential investors from learning more about the company, which is why it is driving funding down in the case of Gilead Sciences.

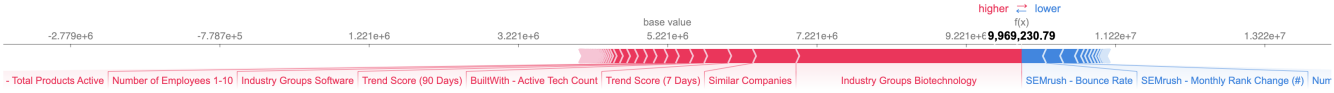

Figure 16: The SHAP force plot for Gilead Sciences

Parting Pro Parting Pro is a small-sized healthcare startup headquartered in Los Angeles, California that focuses on creating the best death care experience for families and professionals. We observe in figure 17 that the main factors driving funding raised lower are the `Number of Contacts` and the Bounce Rate. Upon a deeper dive into Parting Pro's Crunchbase profile, we see that the company has 0 contacts listed. This is a negative factor in influencing the investment decisions made by private equity investors as it makes it more difficult for potential investors to get in touch with and learn more about the company. Furthermore, the high `Bounce rate`, as examined in the case of Gilead Sciences, is also a negative factor for similar reasons.

Figure 17: The SHAP force plot for Parting Pro

Trestle Biotherapeutics Trestle Biotherapeutics is a small-sized heathcare startup headquartered in San Diego, California that focuses on developing bio-engineered kidneys for patients with kidney diseases. We observe in figure 18 that the primary factors driving funding raised lower are the `Number of Contacts` and the `90-day Trend Score`. Similar to the case with Parting Pro, a deeper dive into Trestle Biotherapeutics' Crunchbase profile reveals that the company has 0 contacts listed, which, as examined earlier, can negatively influence investment decisions made by potential private equity investors. Furthermore, the `90-day Trend Score` of the company has been negative on average since it was founded in 2020. This has negative implications for potential investors as it is an indicator of the company's diminishing popularity.

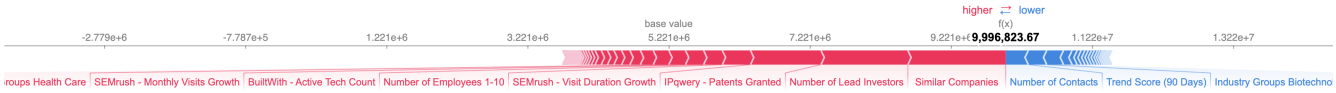

Figure 18: The SHAP force plot for Trestle Biotherapeutics

7 Discussion

The approach proposed and followed in this paper has several advantages:

1. **Data**: We make use of highly recent data available on Crunchbase for more up-to-date results.
2. **Pre-processing**: We make use of the most appropriate feature encoding techniques to get the most out of our features.
3. **ML**: We make use of Gradient Boosted Decision Trees, which is a tree-based ML model for our analysis. Not only does the ML model achieve an extremely high performance on our dataset, thus, ensuring a high degree of faithfulness in our analysis, it also allows for a human-interpretable analysis of the features driving the predictions it makes.
4. **SHAP**: We make use of SHAP as our ML explainability method in our analysis. SHAP enforces the assumption of independence of features used in the ML model during the computation of feature importance scores to ensure that the predictions made are a result of causal inference, thereby preventing counter-intuitive explanations from arising out of correlations among input features or between any of the input features and the output feature [28].

On the other hand, our approach also suffers from several disadvantages. Primarily, we lack several important features and their characteristics that could have been incorporated into the ML model for a more comprehensive analysis. For instance, the `Number of Articles` feature can lead to either a positive or negative influence on the decisions made by private equity investors depending on whether or not these articles about the company were positive or negative. Perhaps for future work, a Sentiment Analysis model [42] could be used to predict whether or not these articles were positive or negative and incorporated into the model for improved performance. Another similar example is the lack of information on who the investors in the companies were. As mentioned earlier, the paper by Li *et al.*[32] found that the credibility of lead investors directly impacts the funding a company raises because lead investors with better track-records attract more followers, which in turn, leads to higher funding received by these companies. This is another direction that can be pursued as future work for this paper. Perhaps another ML model can be used to predict the credibility score of investors based on their track records and incorporated into the model for improved performance.

8 Conclusion

In this paper, we have successfully proposed, implemented, and executed a novel approach to evaluating the investment decisions made by private equity investors in seed-stage healthcare startups. The approach leverages highly recent investment data from Crunchbase to train a Gradient Boosted Decision Tree ML model that achieves an impressive performance on the same dataset. SHapley Additive exPlanations, an ML explainability method is then applied to the ML model in order to probe it and gain insights into the investment patterns made by private equity investors. Finally, we successfully made use of our approach at both a global scale (top and bottom 100 companies) and a local scale (top and bottom 3 companies) to delve deeper into the factors driving these investments and showed that the faithful ML explanations highly corroborate with results found in literature.

References

[1] B. Cerullo and B. Sommer, "Helping healthcare entrepreneurs: a case study of angel healthcare investors, llc," *Venture Capital: an international journal of entrepreneurial finance*, vol. 4, no. 4, pp. 325–330, 2002.

[2] A. Durai, B. Li, S. Metkar, M. Pelayo, and N. Phillips, "Challenges in a biotech startup healthcare nuts," *Kellog School of Management Northwestern University*, 2006.

[3] P. Lehoux, F. Miller, and G. Daudelin, "How does venture capital operate in medical innovation?," *BMJ innovations*, vol. 2, no. 3, 2016.

[4] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *Ieee Access*, vol. 8, pp. 42200–42216, 2020.

[5] C. Ünal, "Searching for a unicorn: A machine learning approach towards startup success prediction," Master's thesis, Humboldt-Universität zu Berlin, 2019.

[6] D. Dellermann, N. Lipusch, P. Ebel, K. M. Popp, and J. M. Leimeister, "Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method," *arXiv preprint arXiv:2105.03360*, 2021.

[7] G. Ross, S. Das, D. Sciro, and H. Raza, "Capitalvx: A machine learning model for startup selection and exit prediction," *The Journal of Finance and Data Science*, vol. 7, pp. 94–114, 2021.

[8] V. Shrivastava, "Predicting and modeling performance of venture capitalists by using linkedin and machine learning,"

[9] V. Wu and C. Gnanasambandam, "A machine-learning approach to venture capital," *McKinsey Quarterly*, vol. 27, 2017.

[10] J.-M. Dalle, M. Den Besten, and C. Menon, "Using crunchbase for economic and managerial research," 2017.

[11] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," *arXiv preprint arXiv:1604.06737*, 2016.

[12] A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[13] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1310–1315, Ieee, 2016.

[14] L. Torgo and J. Gama, "Regression by classification," in *Brazilian symposium on artificial intelligence*, pp. 51–60, Springer, 1996.

[15] D. M. Allen, "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, vol. 13, no. 3, pp. 469–475, 1971.

[16] D. Wallach and B. Goffinet, "Mean squared error of prediction as a criterion for evaluating and comparing system models," *Ecological modelling*, vol. 44, no. 3-4, pp. 299–306, 1989.

[17] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[18] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[19] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[20] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] S. Suthaharan, "Decision tree learning," in *Machine Learning Models and Algorithms for Big Data Classification*, pp. 237–269, Springer, 2016.

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[24] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[25] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[27] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *International conference on machine learning*, pp. 9269–9278, PMLR, 2020.

[28] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen, "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models," *Advances in neural information processing systems*, vol. 33, pp. 4778–4789, 2020.

[29] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable ai: A causal problem," in *International Conference on artificial intelligence and statistics*, pp. 2907–2916, PMLR, 2020.

[30] I. Covert, S. M. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17212–17223, 2020.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[32] X. Li, Y. Tang, N. Yang, R. Ren, H. Zheng, and H. Zhou, "The value of information disclosure and lead investor in equity-based crowdfunding: An exploratory empirical study," *Nankai Business Review International*, 2016.

[33] K. Babs, "Semrush review: Best competitor keyword research tool," 2012.

[34] B. Chitkara and S. M. J. Mahmood, "Importance of web analytics for the success of a startup business," in *International Conference on Recent Developments in Science, Engineering and Technology*, pp. 366–380, Springer, 2019.

[35] D. Banerji and T. Reimer, "Startup founders and their linkedin connections: Are well-connected entrepreneurs more successful?," *Computers in Human Behavior*, vol. 90, pp. 46–52, 2019.

[36] R. Nanda and M. Rhodes-Kropf, "Investment cycles and startup innovation," *Journal of Financial Economics*, vol. 110, no. 2, pp. 403–418, 2013.

[37] G. Foster and T. Davila, "Startup firms growth, management control systems adoption, and performance," *Management Control Systems Adoption, and Performance (July 2005)*, 2005.

[38] D. Klein Velderman, "Ipos and syndication, a networking effort? how different network structures change the importance of previous investment experience of venture capital firms," 2021.

[39] A. Conti, J. Thursby, and M. Thursby, "Patents as signals for startup financing," *The Journal of Industrial Economics*, vol. 61, no. 3, pp. 592–622, 2013.

[40] S. Kivisaari, R. Lovio, E. Väyrynen, *et al.*, "Managing experiments for transition. examples of societal embedding in energy and health care sectors," *System innovation and the transition to sustainability: Theory, evidence and policy*, pp. 223–250, 2004.

[41] J. Adler-Milstein, D. W. Bates, and A. K. Jha, "A survey of health information exchange organizations in the united states: implications for meaningful use," *Annals of internal medicine*, vol. 154, no. 10, pp. 666–671, 2011.

[42] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.