Startup Success Prediction and Robust Portfolio Allocation

Aziz Malouche
Daniel Rapoport
Maxime Basse

15.095 Machine Learning Under a Modern Optimization Lens
December 7, 2024

1 Problem Description

Predicting startup success and optimizing investment decisions present significant challenges due to the volatile nature of outcomes and limited available data. This research proposes a novel two-step approach to address these issues:

1. **Prediction:** We employ advanced multi-modal data fusion techniques, combining tabular and textual data to enhance predictive accuracy. Our method leverages transfer learning and TabText [1] to maximize the value of unstructured and categorical data, respectively extending similar past approaches [2].
2. **Prescription:** Building on these predictions, we explore robust optimization frameworks for portfolio management. We compare three approaches: point prediction, robust point prediction, and weighted average methods to determine the most effective strategy for investment decision-making.

This integrated approach aims to bridge the gap between startup success prediction and practical investment strategies, offering a more comprehensive solution than current models that often focus solely on prediction.

2 Data

2.1 Raw Data

The data was collected from Crunchbase, a platform that provides comprehensive information on companies, their funding rounds, and acquisition events. Using access provided by MIT Sloan, we accessed Crunchbase's API to gather descriptive features about startups. Due to limitations of data-access through the API we were forces to manually scrape financial data for over 10,000 start-ups from the Crunchbase web-interface. The scope of this analysis was limited to U.S.-based startups only that were founded between 2008 and 2019 and had successfully received Series B funding. This selection ensures that our analysis centers around mid-maturity startups, which are of particular interest to venture capitalists (VCs), yet provide funding history and richer data to enhance comparability for our prediction and prescription. Choosing 2019 as the cut-off year ensures that every startup in our dataset has at least five years of historical data after reaching Series B. This five-year time horizon is essential for defining and evaluating our success metric, as will be detailed later in this report. The descriptive features collected for each startup include the number of founders, founding date (6), industry group (e.g., AI, biotechnology - 3), and headquarters region (e.g., Bay Area, Greater Boston Area - 5).

2.2 Feature Engineering

Our feature engineering approach systematically categorized startup data into three key sections: temporal features, funding-related features, and company descriptive features. This structured methodology allowed us to capture both granular and aggregated insights while addressing the varying paths startups take to Series B funding. For example, some startups may have multiple Series A rounds or varying numbers of seed rounds. To account for these differences, we developed a methodology that aggregates funding events while retaining granular detail, ensuring robust and comparable feature representation.

Temporal Features Temporal features capture the time intervals and milestones in a startup's history, offering insights into the timing and pace of its growth.

Funding-Related Features Funding-related features detail the financial history and magnitude of funding rounds, providing a measure of a startup's capital acquisition and investor confidence.

Company Descriptive Features Company descriptive features capture static attributes and descriptions that provide contextual information about a startup's market and operations.

Temporal Features
Days Between Founding and Earliest Funding Round
Days Between Founding and First Seed/Pre-Seed
Days Between Founding and Last Seed/Pre-Seed
Days Between Founding and Series A
Days Between Founding and Series B

Table 1: Summary of Temporal Features

Funding-Related Features
Earliest Funding Round – Type
Earliest Funding Round – Money Raised (in USD)
Series A Money Raised (in USD)
Total Money Raised Before Series B
Total Funding Rounds Before Series B
Number of Seed/Pre-Seed Rounds

Table 2: Summary of Funding-Related Features

Company Descriptive Features
Number of Founders
Industry Group (e.g., AI, Biotechnology)
Headquarters Region (e.g., Bay Area, Greater Boston Area)
Textual Description of Startup (used for embeddings)

Table 3: Summary of Company Descriptive Features

2.2.1 Defining Success

Defining the success of a startup is inherently challenging, as it can be viewed from various perspectives. From our review of the literature, success is often estimated in a binary fashion, where the occurrence of an acquisition, a significant funding round, or an IPO is considered a success [3]. This binary definition aligns with the common trajectory of startups, which either succeed or perish in most cases.

While this binary metric is reasonable, we wanted to explore an additional, more nuanced measure of success that provides a detailed account of the extent of a startup's achievements in the case of positive outcomes. Specifically, we focused on the growth in valuation as a continuous measure of success. The valuation growth over the five years following a Series B round serves as a strong indicator of the progress and growth that a startup has managed to achieve.

Therefore, in the remainder of this report, we adopt a dual definition of success:

- **Binary Success:** A startup is deemed successful if it achieves at least one significant event (e.g., acquisition, IPO, or a large funding round) in the five years following its Series B round.
- **Continuous Success:** The success is measured by the valuation growth multiple over the five-year period, comparing the valuation just before the Series B round to the valuation five years afterward.

This dual approach provides a more comprehensive understanding of startup success, combining both qualitative milestones and quantitative growth.

3 Methods

3.1 Predictive Framework

For binary success classification, we employ a maximum likelihood approach to estimate the probability of success. For the continuous variable, representing valuation growth, we minimize the mean squared error (MSE) to capture the extent of a startup's achievements. To enhance predictive performance continuous growth was log-transformed prior to prediction-model execution. Exemplary for the classification part, the problem at hand can be formalized as follows:

Given a dataset $S_N = \{(x^1, y^1), \ldots, (x^N, y^N)\}$, where:

- x^i: Feature vector of the i-th startup, comprising structured (numerical/categorical) and unstructured (text) features.
- $y^i \in \{0, 1\}$: Binary label indicating startup success ($y^i = 1$) or failure ($y^i = 0$).

We aim to learn a model $f : X \to [0, 1]$ such that $f(x^i) = \hat{y}^i$, predicting the probability of success. The model is trained by solving:

$$\hat{f} \in \underset{f \in \mathcal{F}}{\arg\min} \; -\frac{1}{N} \sum_{i=1}^{N} \left(y^i \log f(x^i) + (1 - y^i) \log(1 - f(x^i)) \right) + \lambda \Omega(f),$$

where $\ell(f(x^i), y^i)$ is the binary cross-entropy loss, $\mathcal{F}$ is the hypothesis space (in this case Decision Tree, Random Forest, XGBoost, LightGBM, CatBoost and Optimal Classification Trees), $\Omega(f)$ is a regularization term, and λ controls regularization.

Modeling Approach. Initial efforts focused on traditional prediction models, including CART, Random Forests, and XGBoost, LightGBM and CatBoost. These models were applied to predict the binary success metric, starting with tabular/numerical data and later incorporating categorical data, such as the industry or location of a startup. Lastly, an Optimal Classification Tree was also used for the binary success metric, and an Optimal Regression Tree for the continuous success metric.

TabText & Embeddings. To enhance the predictive power of categorical features, we employed TabText, a method that transforms tabular data into textual representations to extract contextual embeddings. For categorical variables, such as industry and location, meaningful sentence prefixes were appended, and embeddings were generated using BERT. For unstructured textual descriptions, we initially utilized BERT, followed by Llama for richer contextual understanding. To reduce embedding dimensionality and improve computational efficiency, two techniques were applied: PCA and a headless feedforward neural network (FFNN). The FFNN was trained on the original embeddings and the dependent variable, with its intermediate outputs used as reduced-dimension embeddings. Finally, all embeddings were concatenated with structured data and evaluated across the models mentioned above.

Validation. A stratified 5-fold cross-validation procedure was applied to evaluate all models, maintaining class balance across folds. Hyperparameter tuning for CART, RandomForest, XGBoost and LightGBM was conducted using randomized search, optimizing AUC-ROC using the train- and validation set only, while the test-set was preserved for best-model execution as a last step.

Evaluation. Model performance was primarily assessed using AUC-ROC for binary classification, measuring the trade-off between true positives and false positives. While models like XGBoost and LightGBM offered higher predictive accuracy, Optimal Classification Trees provided interpretability, highlighting key success factors. For continuous metric prediction R^2 was utilized as the primary evaluation metric.

3.2 Prescription

Once good predictions of startup success are achieved, turning those predictions into a balanced and secure investment portfolio remains a significant challenge. A default approach consists of investing in the startup for which we predict the largest continuous success. This approach, referred to as **Point Prediction**, directly uses the predictions from our machine learning models in the optimization formulation without accounting for uncertainty. The Point Prediction formulation can be expressed as follows:

$$\max_{z} \sum_{i=1}^{N} \hat{y}(x^i)(\hat{r}(x^i) - 1)z_i, \quad \text{s.t.} \quad \sum_{i=1}^{N} z_i \leq B, \quad z_i \geq 0 \quad \forall i.$$

This formulation is trying to maximize the expected profit under a budget constraint. The binary and continuous success are estimators obtained with machine learning predictions using the startup features x^i. While straightforward, this method can be overly optimistic and risky if the predictions are not reliable. To account for uncertainty in predictions, we explored two alternative approaches:

1. **Robust Formulation:** In this approach, we introduce an uncertainty set to handle potential errors in predictions. Specifically, we allow up to $\Gamma\%$ of the predicted success labels to be flipped, simulating a worst-case adversarial scenario. The intuition behind this formulation is that it forces the optimization to diversify the investment portfolio. By avoiding an "all-in" approach on one or a few startups, the model guards against scenarios where adversarial flips in predicted success labels could lead to zero returns. The uncertainty set is defined as:

$$\mathcal{U} = \left\{ \Delta y \in \{0,1\}^N \;\middle|\; \frac{1}{N} \sum_{i=1}^{N} \Delta y_i \leq \Gamma \right\}.$$

The robust optimization problem is then expressed as:

$$\max_{z} \min_{\Delta y \in \mathcal{U}} \sum_{i=1}^{N} |\hat{y}(x^i) - \Delta y_i|(\hat{r}(x^i) - 1)z_i,$$

$$\text{s.t.} \quad \sum_{i=1}^{N} z_i \leq B, \quad z_i \geq 0, \quad \Delta y_i \in \{0,1\}, \quad \forall i.$$

Finally, after linearizing the absolute value and relaxing the binary variables to $0 \leq \Delta y_i \leq 1$ the inner minimization problem can be reformulated as:

$$\begin{aligned} \min_{\mathbf{t}, \Delta y} \quad & \sum_{i=1}^{N} t_i(\hat{r}(x^i) - 1)z_i \\ \text{s.t.} \quad & \frac{1}{N} \sum_{i=1}^{N} \Delta y_i \leq \Gamma, \quad \Delta y_i + t_i \geq \hat{y}(x^i), \quad -\Delta y_i + t_i \geq -\hat{y}(x^i) \quad \forall i \\ & \Delta y_i \leq 1, \Delta y_i \geq 0, \quad t_i \geq 0, \quad \forall i. \end{aligned}$$

Taking the dual of the inner linear optimization problem, we obtain a final robust formulation of our portfolio optimisation as:

$$\max_{\theta, u, v, w, z} N\Gamma\theta + \sum_{i=1}^{N} \left(\hat{y}(x^i)(u_i - v_i) + w_i \right),$$

$$\text{s.t.} \quad \sum_{i=1}^{N} z_i \leq B, \quad \theta + u_i - v_i + w_i \leq 0 \, \forall i,$$

$$u_i + v_i \leq (\hat{r}(x^i) - 1)z_i \,\forall i,$$
$$u_i \geq 0 \,\forall i, \quad v_i \geq 0 \,\forall i,$$
$$w_i \leq 0 \,\forall i, \quad z_i \geq 0 \,\forall i.$$

2. **Weighted Average Prescriptions:** This approach leverages predictions from both our Optimal Classification Tree (for binary success) and Optimal Regression Tree (for continuous success). By adopting a stochastic optimization perspective, we account for uncertainty in the binary and continuous success of each startup. Startups are categorized into the leaves of the trees, and the predictions from these leaves on the train set samples are combined to compute a weighted average. The optimization problem can be written as:

$$\max \sum_{i=1}^{N} \left(\frac{1}{|L_1(s_i) \cup L_2(s_i)|} \sum_{u \in L_1(s_i) \cup L_2(s_i)} (r(u) - 1)y(u) \right) z_i,$$
$$\text{s.t.} \quad \sum_{i=1}^{N} z_i \leq B, \quad z_i \geq 0 \quad \forall i.$$

Here, $L_1(s_i)$ and $L_2(s_i)$ represent the sets of predictions from the classification and regression trees for the startups in the portfolio and u are train samples.

These two approaches shall provide more robust and realistic strategies for investment portfolio optimization by accounting for uncertainty in predictions.

3.2.1 Evaluating Prescriptions

To evaluate the performance of each prescriptive method, we simulate the investments prescribed by each strategy and measure the resulting profits over the five years following the Series B investment. A startup that fails yields zero returns, and the entire investment in it is lost. A startup that succeeds experiences valuation growth, which we use to calculate the returns on investment. The profit made from a successful startup is calculated as $(r - 1) \times$ investment, where r is the valuation growth of the startup over the five years following its Series B.

Our trees were trained on an 80% random split of the total dataset, leaving 1,500 startups available for testing. To ensure fairness in evaluation, we designed a benchmark procedure: for a given portfolio size K, we construct $1500/K$ different portfolios, each tested against the three prescriptive strategies. The resulting profits are collected, and two key metrics are extracted for comparison:

- **Median Profit:** Represents the typical performance of a strategy across portfolios, providing a balanced view of long-term investments (e.g., 5 years in this case). Unlike the average return, which can be skewed by rare, highly lucrative events, the median reflects the outcome an investor is more likely to experience.
- **Risk Adjusted Return:** This measures the variability of profit across the different portfolios by taking the expected value of return divided by the standard deviation, providing an indicator of the risk associated with each strategy.

4 Results

4.1 Prediction

Prediction using traditional methods was performed on a binary outcome variable indicating whether a startup is predicted to be successful. Success was defined as achieving any of the following within 5 years after a Series B funding round: an Initial Public Offering (IPO), an acquisition, or a Series C funding round.

The performance of five predictive models was evaluated—**CART, Random Forest, XGBoost, LightGBM, and CatBoost**—using **AUC** as the evaluation metric. The models were assessed under the following conditions:

1. Numeric data as predictors only.
2. Numeric and categorical data as predictors.
3. Numeric and categorical data with **TabText** embeddings fed into **BERT**.
4. Numeric and categorical data with startup descriptions using **BERT**.
5. Numeric and categorical data with startup descriptions using **Llama**.
6. Numeric and categorical data with startup descriptions using **transfer learning**.

Table 4: Comparison of Prediction Models on Binary Success

Model	**Num.**	**Num. + Cat.**	**Num. + Cat. + TabText**	**Num. + Cat. + BERT**	**Num. + Cat. + LLama**	**Num. + Cat. + Transfer**
CART	0.61	0.66	0.61	0.61	0.61	0.63
Random Forest	0.75	0.78	0.73	0.72	0.72	0.76
XGBoost	0.75	0.79	0.77	0.75	0.75	0.77
LightGBM	0.76	0.79	0.78	0.77	0.76	0.79
CatBoost	0.78	0.80	0.79	0.78	0.77	0.80
OCT	-	**0.70**	-	-	-	-

Among these models, CatBoost demonstrated the highest performance, achieving the best AUC score of 0.80 when using numeric and categorical features for predictions (Confusion Matrix 2). Neither TabText nor embeddings of the description (both raw as well as lower-dimensional) improved the performance of the model, indicating that for our data sample no additional predictive power was incorporated in those features. CatBoost is a boosting algorithm known for its ability to handle categorical features efficiently and produce robust, high-performance models. Notably, the highest AUC was achieved when using the combination of numeric and categorical data. The inclusion of additional data types, such as tabular text, BERT embeddings, LLama models, or transfer learning techniques, did not enhance the model's performance and may have introduced noise into the predictions, making some worse than how they were without it.

An **Optimal Classification Tree (OCT)** [4] was trained using only tabular (numeric & categorical) data. The optimal classification tree is highly interpretable and achieves an AUC of 0.70. While this is lower than some traditional models, it still demonstrates strong performance.

Analyzing the tree provides valuable insights for investors, revealing that the most important features for predicting startup success. The optimal classification tree identified key predictors of startup success, including the number of funding rounds, days between funding rounds and the Series B funding round, and days between founding and the first seed or pre-seed round. These insights suggest that startup success can be predicted by the speed at which a startup begins to raise funds. This is logical, as early and rapid fundraising indicates that other investors have confidence in the startup's potential for growth.

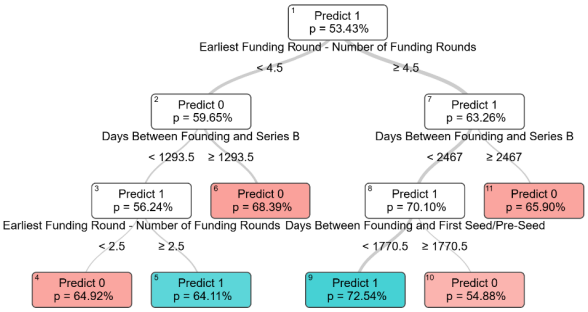

Figure 1: Optimal Classification Tree

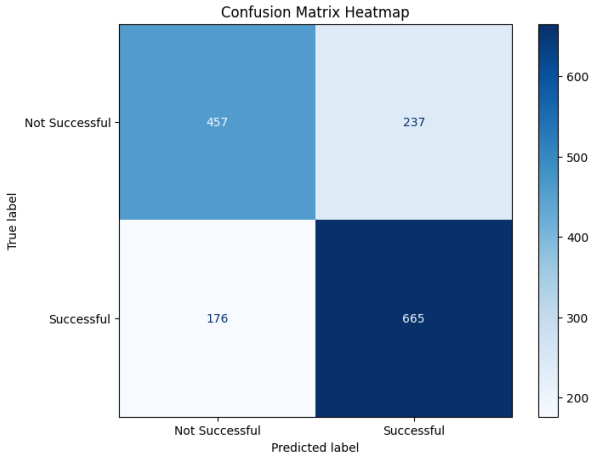

Figure 2: Confusion Matrix for Binary Classification (LightGBM - Numerical & Categorical Features)

Next, the **Optimal Regression Tree (ORT)** was used to predict the continuous variable representing predicted valuation growth multiples. The Optimal Regression Tree is interpretable, but due to its larger size, it is more challenging to interpret than the Optimal Classification Tree. It achieves an R^2 of 0.46, which is expected, as predicting a continuous variable is inherently more difficult than predicting a binary outcome. This tree can be found in the appendix.

We decide to utilise both the **OCT** as well as the **ORT** for the prescriptive task, as they deliver acceptable performance, while providing clear definitions of prediction-neighborhoods utilized in the weighted average approach.

4.2 Prescription

The table below summarizes the performance of the three prescriptive methods—Point Prediction, Weighted Average, and Robust Point Prediction—across different portfolio sizes (K) in terms of Median Return and Risk-Adjusted Return. This example assumes investors have $1000 to allocate, with the results under 'Median Return' expressed in thousands for the median expected return.

Table 5: Comparison of Prescriptive Methods in Terms of Median Return and Risk-Adjusted Return for Different Portfolio Sizes (K)

Portfolio Size	**Median Return**			**Risk-Adjusted Return**		
	Point Pred.	**Weighted Avg.**	**Robust Point Pred.**	**Point Pred.**	**Weighted Avg.**	**Robust Point Pred.**
10	9.6	**11.5**	5.8	0.27	**0.29**	0.24
25	15.7	**15.9**	9.4	0.30	0.29	**0.43**
50	16.6	**18.2**	17.2	0.36	0.37	**0.97**
100	3.0	20.8	**23.7**	0.29	0.34	**1.13**

For a portfolio size of 10, for example, and given a test set of 1500 startups, we would create 150 different portfolios, each consisting of 10 startups randomly assigned. Point prediction, weighted average, and robust point prediction would then be applied to each portfolio, with the option to select in between 1-10 of the startups and can select how much to invest in each. To mitigate the impact of outliers, the median prediction across all portfolios was calculated for each model. This approach provides a more comprehensive evaluation framework compared to testing on a single test set of 1500 startups. Subsequently, the risk-adjusted return

was calculated by taking the mean of the expected value from each model and dividing it by the standard deviation. In the previous example, this would involve dividing the average expected return across the 150 portfolios by the standard deviation. This calculation helps assess the stability of investments.

The Weighted Average method yields the highest return for portfolio sizes of 10, 25, and 50, while Robust Point Prediction provides the highest return for a portfolio size of 100 startups. This highlights the strengths of both models under different investment scenarios. When evaluating risk-adjusted return, Robust Point Prediction generally outperforms the other methods. However, both Weighted Average and Robust Point Prediction consistently outperform Point Prediction. The results highlight key differences between the methods. The Robust Point Prediction method was the only one that produced diversified investment portfolios, as it penalizes overly concentrated investments, making it more suitable for scenarios where prediction uncertainty could lead to significant losses. In contrast, the Point Prediction and Weighted Average methods tended to pick only one startup with the highest expected return due to the lack of manual diversification constraints, which is suboptimal in realistic scenarios. This outcome emphasizes the distinct behaviors of the methods under simplified assumptions.

These results illustrate the importance of explicitly incorporating diversification into the optimization problem. In later iterations, we extended our model to include realistic constraints, such as limits on the maximum allocation to a single startup and sector-specific diversification, resulting in point prediction and weighted average investing in multiple startups. This adjustment ensures practical applicability while preserving the advantages of the robust and stochastic optimization approaches.

5 Conclusion

This research integrates predictive machine learning models with prescriptive optimization techniques to address the challenges of startup success prediction and investment decision-making. CatBoost emerged as the best-performing model for binary classification, although textual embeddings did not significantly enhance predictive power for this dataset. Optimal Trees offered better interpretability while achieving comparable AUCs / R^2, making them the preferred choice for final predictions.

On the prescriptive side, the Robust Point Prediction method consistently outperformed in risk-adjusted returns and diversification, while Weighted Average achieved the highest median returns for smaller portfolios. In contrast, Point Prediction's lack of diversification resulted in concentrated and risk-prone investments. These findings emphasize the importance of incorporating diversification and uncertainty handling in portfolio optimization. Future work can enhance this framework by adding dynamic constraints and sector-specific diversification, offering practical tools for data-driven decision-making in venture capital.

6 Work Contributions

Aziz worked on feature engineering, traditional predictive models, optimal trees, and different variations (with / without constraints / robustness) of running the point prediction and weighted average models. Daniel worked primarily on the predictive modeling, incl. embeddings for descriptions and TabText as well as the benchmark implementation for the prescriptive method. Maxime worked on the formulation of the weighted average with leaves from two trees, the robust formulation of that problem, the implementation of the three methods, the benchmark framework and the feature engineering for predictions.
Contributions to the report were shared equally.

7 References

References

[1] Kimberly Villalobos Carballo, Liangyuan Na, Yu Ma, Léonard Boussioux, Cynthia Zeng, Luis R. Soenksen, and Dimitris Bertsimas. *TabText: A Flexible and Contextual Approach to Tabular Data Representation.* arXiv preprint arXiv:2206.10381, 2023. Available at: https://arxiv.org/abs/2206.10381

[2] Abdurahman Maarouf, Stefan Feuerriegel, and Nicolas Pröllochs. *A Fused Large Language Model for Predicting Startup Success.* European Journal of Operational Research, 2024. ISSN: 0377-2217. Available at: https://doi.org/10.1016/j.ejor.2024.09.011. https://www.sciencedirect.com/science/article/pii/S0377221724007136.

[3] J. Arroyo, F. Corea, G. Jimenez-Diaz, and J. A. Recio-Garcia. *Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments.* IEEE Access, vol. 7, pp. 124233–124243, 2019. doi: 10.1109/ACCESS.2019.2938659. Available at: https://ieeexplore.ieee.org/document/8821312.

[4] D. Bertsimas and J. Dunn. *Optimal classification trees.* Machine Learning, vol. 106, pp. 1039–1082, 2017. Springer. Available at: https://doi.org/10.1007/s10994-017-5633-9.

8 Appendix

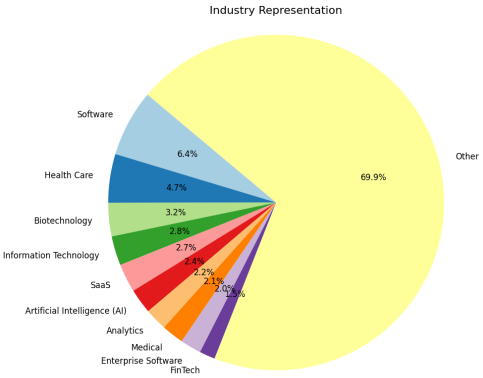

Figure 3: Industry Representation of Startups

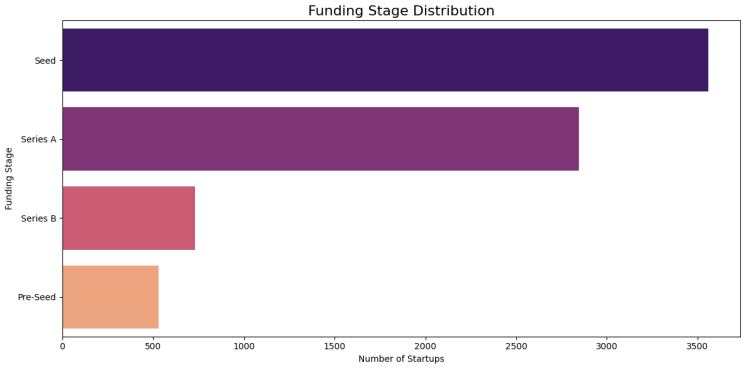

Figure 4: Distribution of Last Funding Round by Year

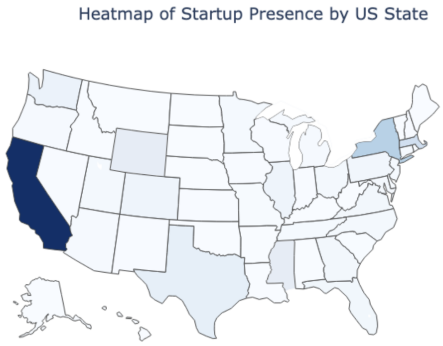

Figure 5: Geographic Distribution of Startups (US Heatmap)

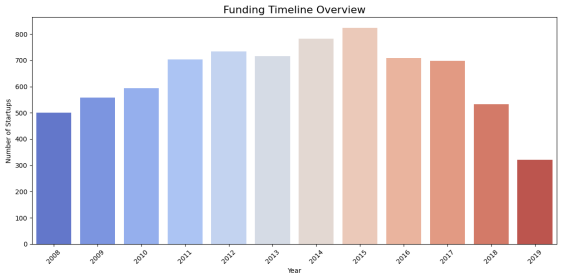

Figure 6: Yearly Distribution of Startup Founding Dates

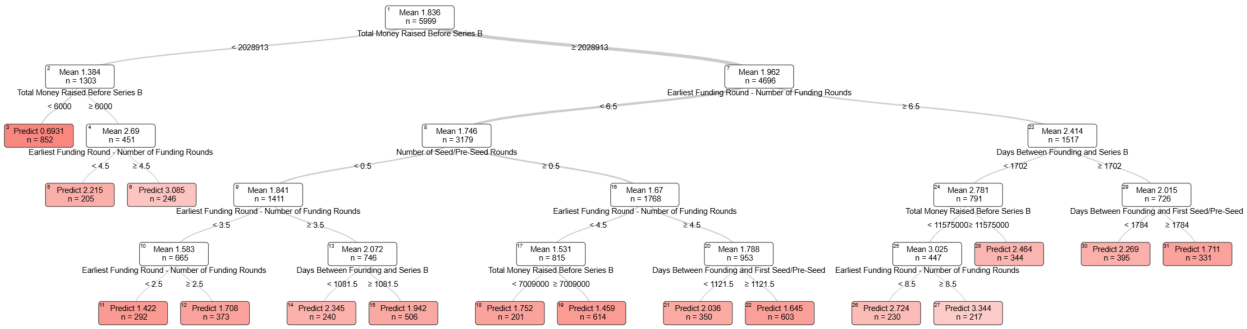

Figure 7: Optimal Regression Tree