Utilizing Public Data and Feature Engineering Techniques to Predict Startup Success

Mansi Achuthan, Nicholas Hausman
Computer Science and Software Engineering, California Polytechnic State University, San Luis Obispo
San Luis Obispo, United States
machutha@calpoly.edu, nrhausma@calpoly.edu

Abstract— **This paper presents an application-oriented approach to predicting startup success by leveraging publicly available data and employing advanced feature engineering techniques. The study extends previous research by incorporating additional data sources, such as Crunchbase, Twitter, and Google Search, and enhancing feature engineering efforts. The feature engineering process covers eight crucial business evaluation categories, including idea viability, team quality, scalability, technology readiness, market presence, competitive landscape, product rollout, and financing status. The prediction model, based on the state-of-the-art CatBoost algorithm, outperforms previous models, providing improved accuracy, precision, recall, and F1-score. The developed prediction pipeline enables real-time funding predictions, assisting potential investors in making informed decisions and increasing the efficiency of startup evaluations. This work contributes to the identification of startups with a higher likelihood of obtaining future funding, using publicly available data and effective feature engineering techniques, ultimately fostering a more data-driven approach to startup investment decisions.**

Keywords— startup, prediction models, funding, publicly available data, screening, feature engineering, methodology, implementation, performance changes, model comparison, takeaways, lessons, project work

I. Introduction

Every year, thousands of startups launch with hopes of revolutionizing the market. However, nine out of every ten startups fail [1], making them risky investments. Nevertheless, the 10% of startups that succeed yield significant profits, especially for their investors. This creates a high risk, high reward environment for potential investors that is difficult to navigate due to lack of access to critical startup data. We seek to inform these investment decisions by utilizing publicly available information in conjunction with feature engineering techniques to identify startups that are likely to raise another round of funding in the future.

Emily Gavrilenko, a Master of Science student in Computer Science at Cal Poly, dedicated her thesis to predicting startup success. Gavrilenko has developed methods for data extraction, feature engineering, and model development, resulting in a prediction model that achieved an F1-score of 0.736 in forecasting startup funding within a three-year time frame [2]. Notably, Gavrilenko's model shows impressive performance compared to existing startup prediction research relying on publicly available data.

Additionally, Gavrilenko and fellow Cal Poly student Kenny Lau began development of a web application designed to screen startups. The website features individual company pages and screening functionality to gain insights into potential startup investments.

We as a team have continued the efforts of Gavrilenko and Lau in gathering and engineering features, creating an improved startup funding prediction model, and improving the Startup Tracker application. This paper describes our efforts to improve this research in detail.

II. Dataset Description and Feature Engineering

For this research, we utilize public internet data to create a dataset. Public internet data comes in many forms, but news, social media, and public financial platforms are categories of sources that contribute to this effort. More specifically, Gavrilenko used Crunchbase, Twitter, and the Google Search APIs to source data to create a predictive model. We considered expanding the breadth of the data sources through platforms such as LinkedIn, Reddit, and some news sources such as TechCrunch. However, difficulty in accessing the data from these sources led us to work within the sources already being used (Crunchbase, Twitter, and Google Search).

From Crunchbase, Twitter, and Google, Gavrilenko engineered over 110 features for each startup company. This section describes our feature engineering efforts, including the business evaluation strategies used in feature creation along with a breakdown of features developed for the different data sources.

A. Business Evaluation

Our team formulated eight business evaluation categories, outlined in Table I, based on discussions with our clients.

TABLE I.

Category	Description
Idea	How sound is the business idea?
Team	Quality of management/founder team?
Scalability	Size of the opportunity?
Tech	Is there a product prototype?
Market	Is there a strategic partnership (sales channels)?
Competitive Environment	Who else is in the space?
Product Rollout	Are there any products yet?
Financing	Any rounds of financing yet? Any need for additional capital?

Table I. The eight business evaluation categories used to select features

These categories serve as a crucial framework for assessing the accuracy of our data in capturing the key factors affecting startup success.

B. Crunchbase

Crunchbase offers a comprehensive database of public and private company information. It provides users with access to profiles of numerous companies, along with information such as general company details, funding and investor information, employee and founder backgrounds, and press references.

In building upon Gavrilenko's Crunchbase feature generation, we conducted a thorough review of the features she sourced from Crunchbase. Table II provides a summary of the 100+ Crunchbase features extracted by Gavrilenko.

TABLE II.

Business Evaluation Category	Example Feature
Idea	Number of Founders
Team	Company Description
Market	Company Industry
Financing	Months Since Last Funding

Table II. A summary of the business evaluation categories covered by Gavrilenko along with an example feature for each category.

We generated more features, focusing on expanding the coverage of the business evaluation categories in the data. A comprehensive list of the Crunchbase features we extracted can be seen in Table III.

TABLE III.

Feature	Description	Category
founder_degree_subjects	A list of unique degree subjects of founder(s) of the company	Team
founder_degree_types	A list of unique degree types of founder(s) of the company	Team
founder_degree_schools	A list of unique schools that founder(s) attended	Team
founder_degree_completed	The proportion of founders that completed their degree(s)	Team
company_name_sentiment	The general attitude towards a company name	Idea
article_title_sentiment	The general tone of news article titles about a company	Idea
avg_article_title_length	The average length of news article titles regarding a company	Idea
hq_greater_region	The region in which the company HQ is located (San Francisco Bay Area, Greater Miami Area, etc.)	Market
similar_company_count	The number of similar companies obtained using Search API	Competitive Environment
similar_company_funding_rounds	The number of similar companies which have been funded in the time frame	Competitive Environment
similar_company_amount_raised	The amount of money raised by the similar companies	Competitive Environment
org_investors	The number of orgs that have invested in the company in a given time frame	Financing
individual_investors	The number of individuals that have invested in the company in a given time frame	Financing
micro_vc_investment_count	The number of micro-VCs that have invested in the company in a given time frame	Financing
vc_investment_count	The number of VCs that have invested in the company in a given time frame	Financing
PE_investment_count	The number of PE that have invested in the company in a given time frame	Financing
HF_investment_count	The number of HF that have invested in the company in a given time frame	Financing
angel_investment_count	The number of angel investors that have invested in the company in a given time frame	Financing

Table III. A list of features generated by our team using data from Crunchbase.

We expanded the scope of business evaluation categories beyond the four covered by Gavrilenko, incorporating the competitive environment category. Furthermore, we introduced additional features that concentrate on founder and investor information, which were not included in Gavrilenko's data collection. This expanded coverage significantly enhances the comprehensiveness of our feature set, thereby enriching the data available for our predictive model.

C. Twitter

Twitter is a well-known social media platform where users can share brief media content, primarily in the form of text-based "tweets". In our project, Twitter serves as a valuable source of public opinion on startups, as well as other relevant business-related information. Within Gavrilenko's data extraction pipeline, tweets are collected if they are posted by a company, are in response to a company's tweet, or reference a company through a related link (e.g., website).

Gavrilenko generated over 60 features from Twitter data within various business evaluation categories. Examples of these features and their corresponding categories can be found in Table IV.

TABLE IV.

Business Evaluation Category	Example Feature
Team	Tweets about Management Changes
Scalability	Average Likes Per Tweet
Market	Tweets about Geographical Expansion
Product Rollout	Tweets about New Products

Table IV. A summary of the business evaluation categories covered by Gavrilenko within the Twitter features along with an example feature for each category.

After assessing the existing coverage of business categories and features, our focus shifted towards expanding the feature list. We carefully examined possible API calls in the Twitter API documentation and reviewed generation pipelines to create the features presented in Table V.

TABLE V.

Feature	Description	Category
influencer_count	Number of tweets by users deemed to be influencers	Market
influencer_sentiment	Average sentiment of tweets by users deemed to be influencers	Market
influencer_avg_retweet_count	Average retweets of tweets by users considered to be influencers	Market
hashtag_sentiment	Sentiment of hashtags in tweets	Market
avg_num_attachments	Average number of tweet attachments	Market
avg_like_monthly_change	Average change of total tweet likes per month	Scalability
avg_likes_competitor	Average tweet likes of similar companies	Competitive Environment
total_likes_competitor	Total tweet likes of similar companies	Competitive Environment
total_tweets_competitors	Total tweets of similar companies	Competitive Environment

Table V. A list of features brainstormed by our team using data from Twitter.

These additional Twitter-based features enhance Gavrilenko's coverage of existing business categories and address the competitive environment category that was not present in the feature set. However, we were unable to implement these features due to the changes with the Twitter API.

D. Additional Sources

Beyond Crunchbase and Twitter, we explored new sources, including Reddit, LinkedIn, and full text news sources like Techmeme and TechCrunch.

Reddit offers rich features that can add value to the model. These include the number of search results for a company name, the count of search results within the r/startups subreddit, and the average upvotes of posts related to a company. Unfortunately, our intended use of Reddit as a data source was impacted by recent changes to their API pricing. If the situation changes in the future, our team has brainstormed a few feature ideas that could be explored.

LinkedIn also presents many potential features that can enhance the model. Examples include the number of company followers, founder followers, and job postings. Given its professional and business-oriented nature, integrating LinkedIn into the feature extraction pipeline would be beneficial. However, the LinkedIn API provides mostly irrelevant data for our research and obtaining data through web scraping user data violates LinkedIn's terms of use. Therefore, interaction with the platform is constrained and uncertain. However, this would be a source of high priority to explore in the future, as we believe some web scraping of non-user (such as company) data could be done.

In addition, full text news sources such as TechCrunch and Techmeme could provide useful features for the modeling effort. While web scraping would be most desirable, paywalls and varying data formats complicate data extraction from online news sources. Nonetheless, platforms like Techmeme and TechCrunch allow for web scraping of full-text content without encountering paywalls, presenting opportunities for extracting interesting features. Potential ideas include average publisher and author rankings on Techmeme, sentiment analysis of article text (on both platforms), and identification of product mentions within articles (on both platforms).

III. Conclusions and Accomplishments

We have contributed to multiple aspects of the project during our work. This section goes over the efforts of our work and the current state of each deliverable of the project.

A. *Feature List*

We developed a feature list to track and analyze the feature set used in the model development process. All features are listed with their names, data sources, data types, business evaluation categories, and more. We believe this list allows stakeholders to offer insights into the further expansion of the feature set along with possible restriction or elimination of unnecessary features. If anyone were to increase the number of data sources or features, this list will allow for those expansions to be both tracked and categorized.

B. *Model Development*

Using the larger feature set, we developed a new prediction model to assess the probability of a company securing funding within the upcoming three years. We decided to focus on this period because it is the average period used by Gavrilenko during model development, making it easier to compare our work to Gavrilenko's. Three years is also a very reasonable window for investment opportunities.

We decided to continue using the CatBoost model, which utilizes gradient boosting on decision trees. Gradient boosting combines multiple decision trees to create a strong predictive model. Each decision tree learns from the mistakes of the previous trees, gradually improving the overall accuracy of the model. The model was developed by Yandex and achieved the best performance of the models utilized by Gavrilenko [3].

The model outperforms the previous model across all the metrics in Table VI.

TABLE VI.

Model	**Metric**	**Score**
Gavrilenko	Accuracy	0.850
	Precision (Macro Avg)	0.821
	Recall (Macro Avg)	0.803
	F1-Score (Macro Avg)	0.811
	Precision (Funded Companies)	0.755
	Recall (Funded Companies)	0.691
	F1-Score (Funded Companies)	0.721
Achuthan and Hausman	Accuracy	0.860
	Precision (Macro Avg)	0.831
	Recall (Macro Avg)	0.809
	F1-Score (Macro Avg)	0.819
	Precision (Funded Companies)	0.773
	Recall (Funded Companies)	0.695
	F1-Score (Funded Companies)	0.732

Table VI. A comparison of scoring metrics observed on the test dataset of Gavrilenko's model versus our model.

While the increases in performance for each individual metric was not large, the F1-score for funded companies increased by 0.01. This metric is the benchmark for research in the field of modeling startups using public data, and Gavrilenko's model had already been the highest performing in this metric among current studies. An increase in this metric signifies a meaningful contribution by the new set of features.

We also analyzed the most notable features in the new model, and some of the new features appear, such as hq_greater_region, num_venture_capital_investors, and founder_degree_schools. See Figure I for the feature importance graph we used in our analysis.

FIGURE I

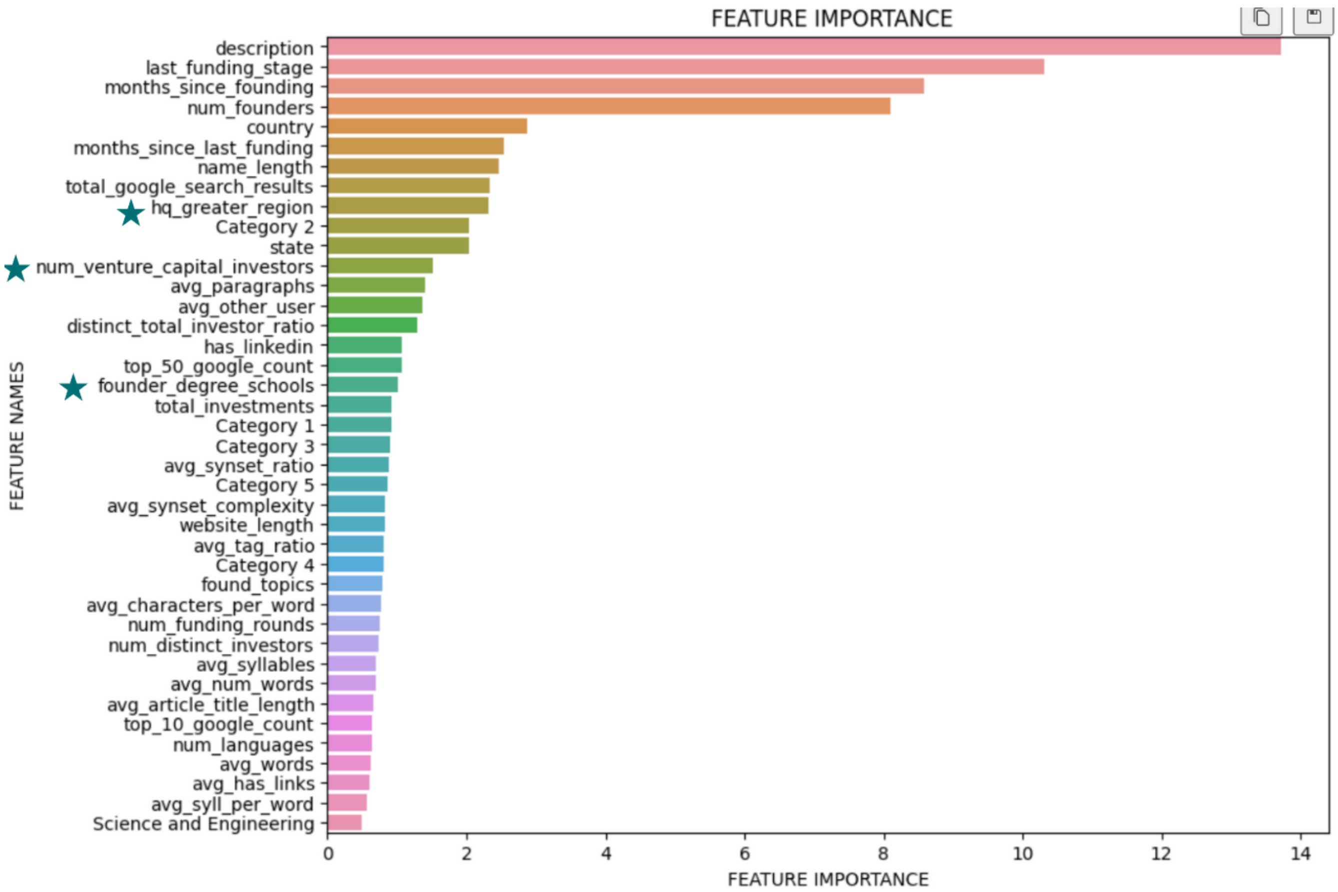


Figure I. The top 40 most significant features in the model. Features are starred if they were created by our team.

C. Prediction Pipeline

Using the prediction model, we created a prediction pipeline to deliver real-time funding predictions on companies. This pipeline supports the end goal of this project in providing investors with live predictions and insights into companies for future investment decisions. The pipeline gathers companies from a Firestore database, hosted on Google Cloud's Firebase, and sends them through data aggregation and feature generation scripts, designed during the feature engineering process. The features are then used by the model to predict funding probabilities. These probabilities, along with the general company data and features we previously aggregated and engineered, are stored in the database.

Within the pipeline, API keys are pulled through a saved file. Because we do not have a Twitter API key, we adjusted the pipe line to only generate features for Twitter if an API key is present. Similarly, the prediction model makes predictions with Twitter features only if the API key is present. This design will allow for simple adjustment if a Twitter API key is acquired.

The current pipeline only makes predictions for a three-year window, hoping to answer the question: "Will company XYZ acquire another stage of funding in the next three years?". The pipeline can easily be adjusted to accommodate other periods once the models are developed and saved.

The pipeline has been running only a pseudo-daily schedule locally but is designed to be deployed to a remote instance where it can be run on a scheduled basis. The pipeline will require a virtual machine instance with significant resources to accommodate the large variety of dependencies and packages involved in the data acquisition and feature generation stages.

D. Startup Tracker Application

The Startup Tracker application is the main deliverable and display of our efforts on the project. The application allows users to research startup companies and gain insights into potential investment opportunities in the private company sector. It provides two main tools to gain insights: the individual company dashboard and the company screener. The application's technology stack consists of a ReactJS frontend, ExpressJS backend, and a Google Firebase database.

Prior to our contributions to the application, the website displayed and utilized the features aggregated from Twitter data. The individual company dashboard showcased time-series graphs for features such as tweet sentiment and number of followers. The screener page listed companies that could be filtered based on the most recently acquired metrics from Twitter, such as follower count and number of tweets.

The dashboard now displays a larger variety of graphics and incorporates data from all integrated feature sources. These include cards for features such as the school the founder(s) attended and degree(s) they obtained. Additionally, a time-series graph of the predicted funding probabilities, acquired from the prediction pipeline, now sits at the top of each company page.

The screener page currently utilizes the entire database, containing over 5,500 companies, by displaying companies with their name, description (extracted from Crunchbase), and an up-to-date prediction value. The page is designed to allow for sorting using any of the displayed columns, such as the prediction score. Additionally, the page can filter companies based on individual feature values, including the Twitter features as aggregated before along with Crunchbase features such as the number of funding rounds acquired.

For an overview of system design for the application, please refer to Figure II.

FIGURE II

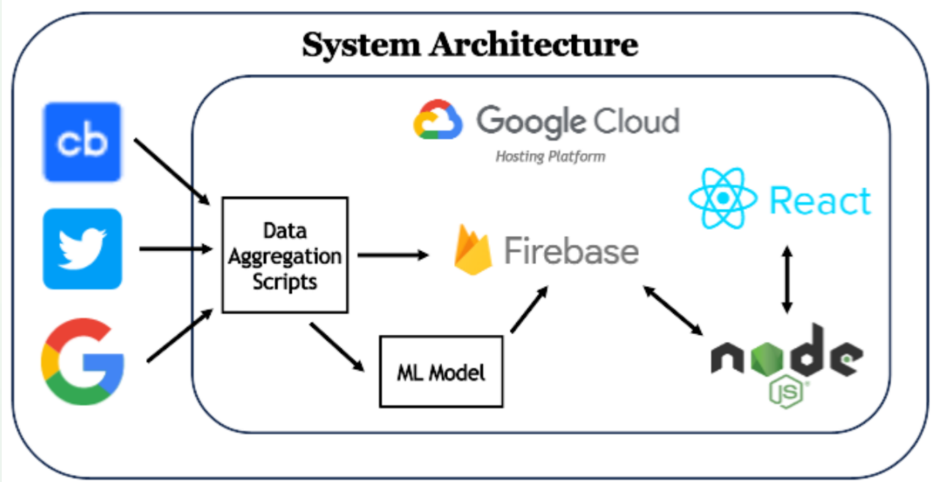


Figure II. The system design of the Startup Tracker Application.

To get an idea of what the application looks like, please refer to Section A of the Appendix.

IV. Future Work

Although significant work has been done on the project as-is, improvements can be made across the board to this work in the future. In this section, we discuss the different aspects of the project that present opportunities for further work.

A. Prediction Pipeline

The most immediate demand for future work on the project would be improvement and deployment of the prediction pipeline. We had hoped to deploy the pipeline as-is to the virtual machine running the application, but the virtual machine did not possess the memory requirements necessary to run the pipeline. The goal would be for the pipeline to be run on a remote instance daily or weekly schedule to provide up-to-date predictions to the application.

B. Data Acquistion and Feature Engineering

The exploration and utilization of more data sources would be useful in improving model performance. We had hoped to explore sources such as LinkedIn, Reddit, and news sources such as TechCrunch during our feature engineering work, but time demands and data access issues prompted us to focus on other work within the project.

In addition to increasing the number of sources, generating more features from the current sources could be explored. The data provided by the Twitter and Crunchbase APIs could easily be used to expand the feature set because of how rich the provided data is. Combining or transforming current features to expand the feature set could be explored as well.

C. Modeling

While we tested different hyperparameters within the CatBoost model, we did not prioritize looking into other model architectures. Incorporating the new feature set and considering the objective of developing a model independent of Twitter, other model architectures such as neural networks could be fit to the dataset and compared with the CatBoost model.

D. Application Development

The Startup Tracker could also be improved from its current form. The UI design of the application is one aspect that can be improved. There are also a variety of features that can be added to the individual company pages. As is, the Google and Crunchbase tabs within the dashboards lack a comprehensive set of displayed features. One example of a useful addition is “Companies you might be interested in”, helping direct users to similar companies. Another idea could be to create a toggle for the prediction graph, allowing users to visualize the trend in prediction scores for models with and without select sources. For example, users could be allowed to visualize a prediction graph for a Twitter-less model, Crunchbase-less model, and more.

V. Acknowledgement

We would like to thank our clients Foaad Khosmood, Mahdi Rastad, and Sadra Amiri for their time and guidance this year. We would also like to thank our professors Alex Dekhtyar and Hunter Glanz for their constant support.

VI. References

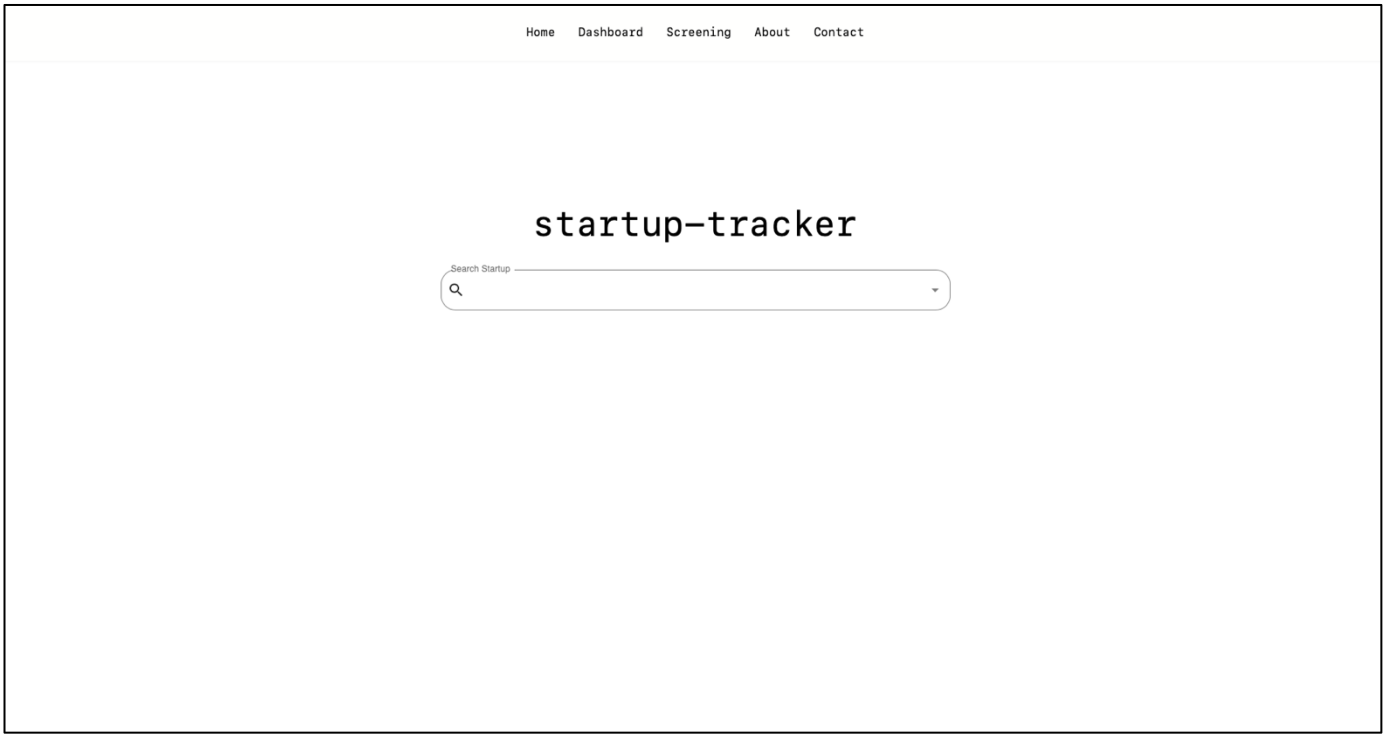
[1] N. Patel, “90% of startups fail: Here’s what you need to know about the 10%,” Forbes, https://www.forbes.com/sites/neilpatel/2015/01/16/90-of-startups-will-fail-heres-what-you-need-to-know-about-the-10/?sh=19ad891b6679. (accessed Jun. 14, 2023).

[2] E. Gavrilenko, “Predicting Startup Success Using Publicly Available Data,” M.S. thesis. Computer Science and Software Engineering, California Polytechnic State University San Luis Obispo, 2022.

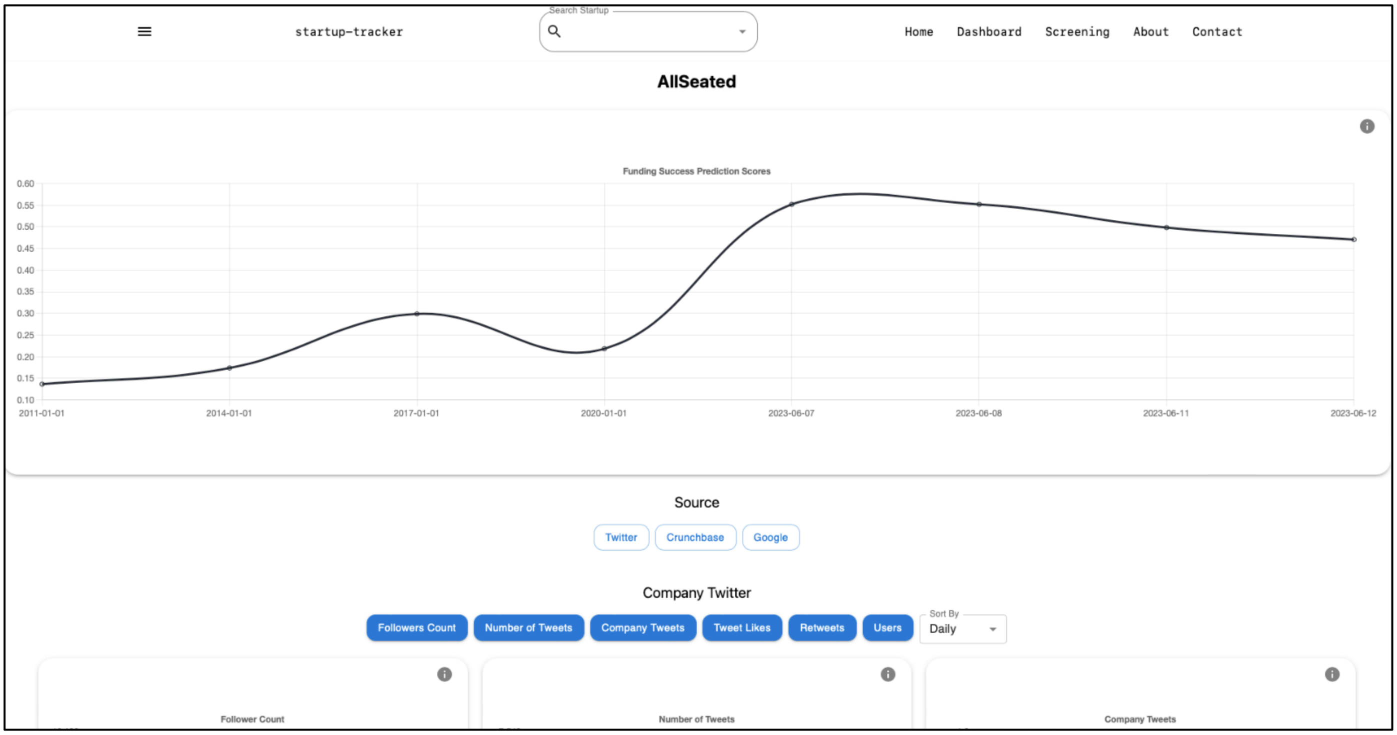
[3] “State-of-the-art open-source Gradient Boosting Library with categorical features support,” CatBoost, https://catboost.ai/#:~:text=CatBoost%20is%20an%20algorithm%20for,CERN%2C%20Cloudflare%2C%20Careem%20taxi. (accessed Jun. 14, 2023).

VII. Appendix

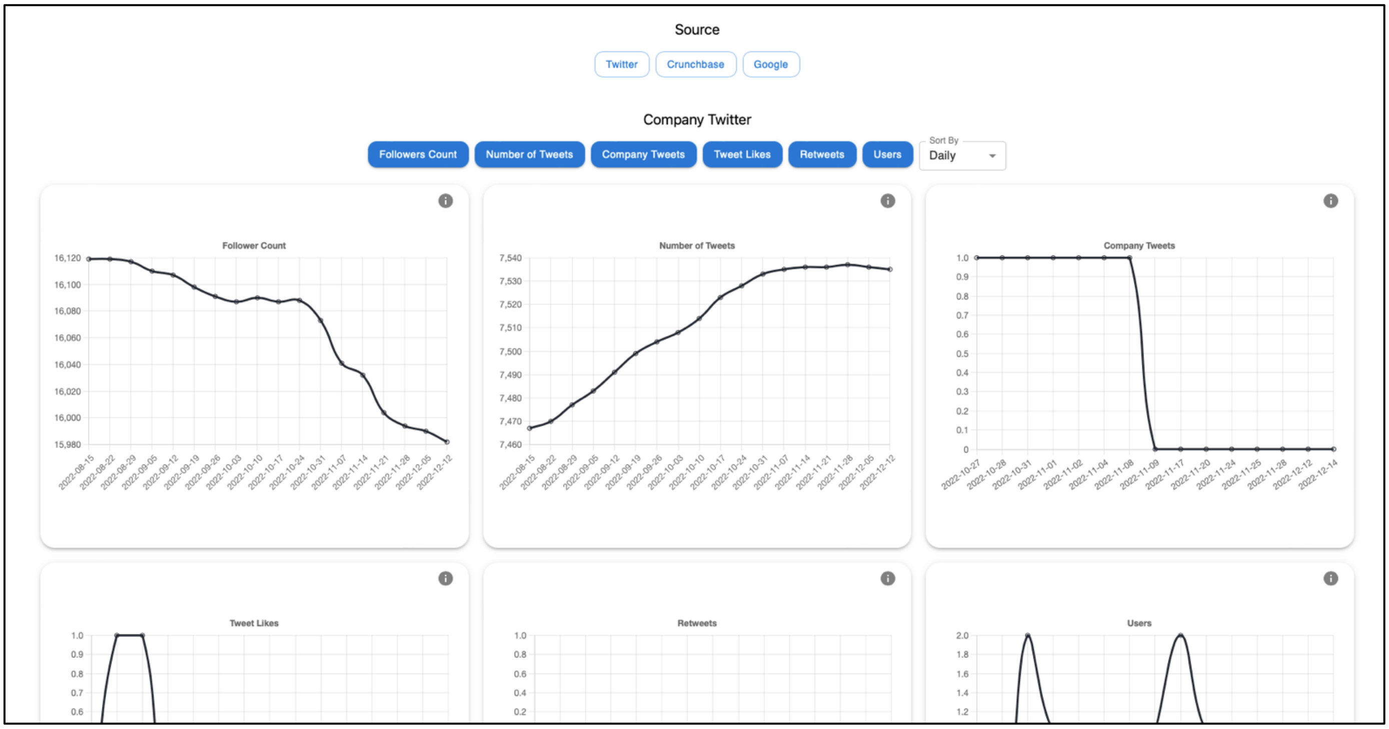
A. Startup Tracker Application



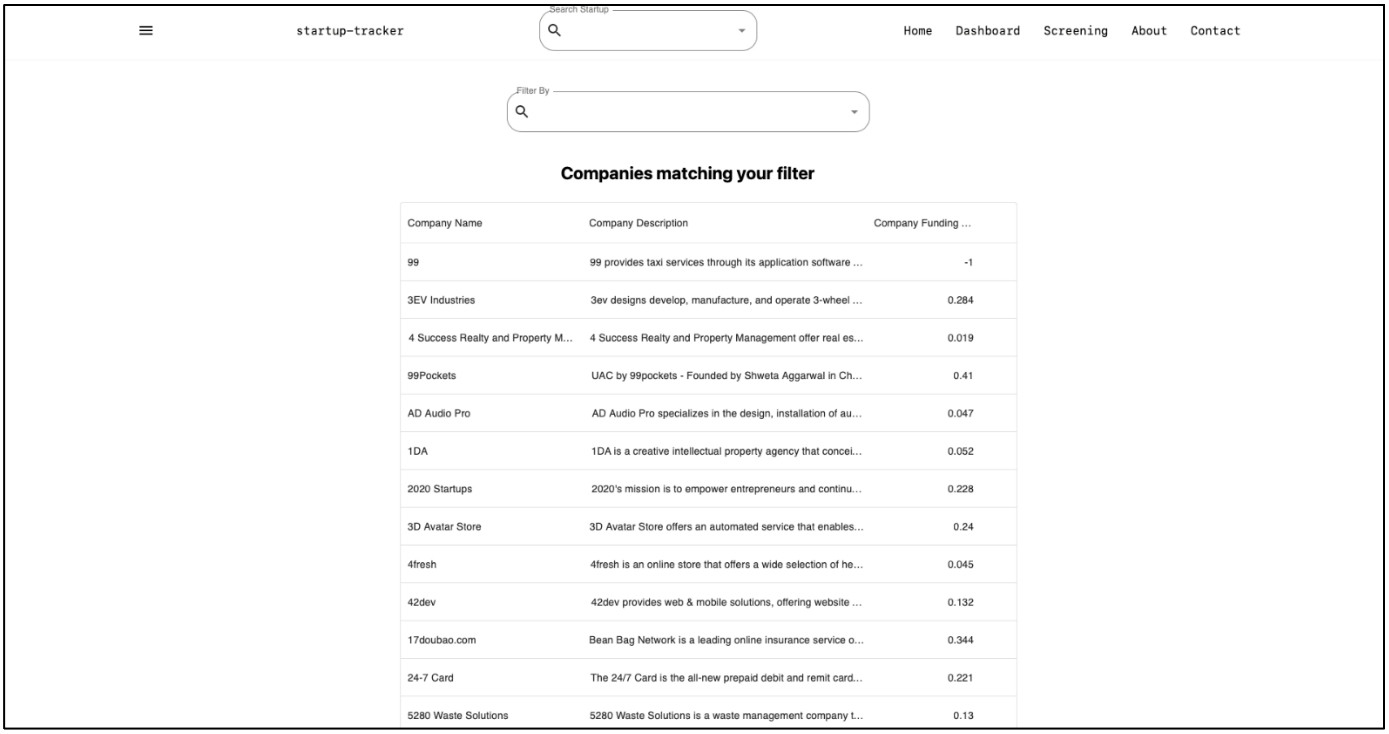
The home page of the Startup Tracker application. Users can navigate to the dashboard and screening from here via the navigation bar at the top. They can also access the company dashboard by searching for a specific company.



Here is an example of what an individual company page would look like in the company dashboard. Each page starts with a Funding Success Prediction graph, visualizing our three-year predictions of funding success over time.



Along with predictions, the company pages showcase aggregated data we pulled from Twitter, Crunchbase, and Google in the data acquisition and feature engineering phase. Each source's visualizations can be turned on and off, allowing users to customize the data to fit their needs.



Companies matching your filter

Company Name	Company Description	Company Funding ...
99	99 provides taxi services through its application software ...	-1
3EV Industries	3ev designs develop, manufacture, and operate 3-wheel ...	0.284
4 Success Realty and Property M...	4 Success Realty and Property Management offer real es...	0.019
99Pockets	UAC by 99pockets - Founded by Shweta Aggarwal in Ch...	0.41
AD Audio Pro	AD Audio Pro specializes in the design, installation of au...	0.047
1DA	1DA is a creative intellectual property agency that concei...	0.052
2020 Startups	2020's mission is to empower entrepreneurs and continu...	0.228
3D Avatar Store	3D Avatar Store offers an automated service that enables...	0.24
4fresh	4fresh is an online store that offers a wide selection of he...	0.045
42dev	42dev provides web & mobile solutions, offering website ...	0.132
17doubao.com	Bean Bag Network is a leading online insurance service o...	0.344
24-7 Card	The 24/7 Card is the all-new prepaid debit and remit card...	0.221
5280 Waste Solutions	5280 Waste Solutions is a waste management company t...	0.13

The screening page, as seen above, by default lists all companies in the database, a description of what they do, and prediction score for how likely it is for the company to raise another round of funding in the next three years. There is also an option to filter by different features such as the number of funding rounds acquired.