Y Combinator Funded Companies: Predicting Success

INFT 6201 - Big Data

Bikram Ghimire(c3502820)
Pradeep Regmi (c3501889)
Rohit Shrestha (c3502400)

Y Combinator

Y Combinator is a startup accelerator program based in Mountain View, California. It was launched in 2005 by Paul Graham, Jessica Livingston, Trevor Blackwell, and Robert Morris.

Here are the key facts about Y Combinator:

• **Early-stage focus** — YC invests in very early-stage companies just starting out. Many companies enter with just an idea or prototype.

• **Huge alumni value** — YC startups have a combined valuation of over $300 billion. This includes unicorns like Airbnb, DoorDash, Stripe, Coinbase, Instacart, and Dropbox.

• **Massive network** — Over 3,000 companies have gone through YC. That creates an invaluable network and pedigree.

• **7% for $125K** — YC invests $125,000 for 7% equity in each company. The standard deal is offered to every startup.

Notable Y Combinator Success Stories

Airbnb — Disruptive hospitality startup now worth close to $100 billion.

Stripe — Online payments leader currently valued at $95 billion.

Cruise Automation — Autonomous vehicle startup acquired by GM for over $1 billion.

DoorDash — Food delivery giant currently worth $57 billion.

Coinbase — The largest US cryptocurrency exchange now worth $54 billion.

Instacart — Grocery delivery app valued at $39 billion.

Reddit — Massively influential social network Reddit.

Datasets: 2024 YCombinator All Companies

Overview:

We're using a dataset from Kaggle, created by Sasha Korovkina , a data scientist at CMI2I,who sourced information from the official Y Combinator companies directory. This dataset includes profiles for all YC-funded companies, capturing essential information like **company name, industry type, geographic location, founder educational background, and founding year**.

This data is critical for answering our core research questions, such as identifying trends in YC's funding focus, spotting growth industries, and understanding the factors contributing to startup success.

Dataset Relevance and Data quality :

The dataset ob

- **Companies:** It includes the data about the company name, slug, company's url, description, team size and status.
- **Founders:** This data gives the information about founder's name, avatar, current company involved in and company slug.
- **Regions:** Contain data about the region, country and address of the startups.
- **Tags:** It provide the information on the area of which the startups are based of like Music, productive, community, Social, AI and Climate etc.
- **Badges:** It is used to distinguish the company success with the use of badges.
- **Industries:** This data define the industry it focus on such as B2B, Consumer, Fintech, Finance and Accounting and AI etc.

Data Quality Issues

- **Lack of Revenue Data:** Doesn’t contain direct success metric to know the revenue growth or market values, user involvement.
- **Lack of History Data:** Doesn’t contain details information about the lunch dates, duration to gain profit, funding process and revenue
- **Limited Data about Team Members:** Only founder data are provided their educational and prior work information but lacks information on the other teams or co-founder information of educational background and prior work experience.
- **Limitation on Data:** This data only provides the company that goes through Y Combinator which lack other successful startup that does not funded by it.

Objective: Why Analyzing Y Combinator-Funded Companies Is Meaningful

Understanding Success Patterns:

Y combinator has funded some of the worlds most successful startups as mentioned earlier. By analysing trends in the Y combinator funded companies, we can have a better insights into the key factors for startup success. Some of the key factors include the industry type, schools where the founders attended, the course they did, their prior company they worked for, team composition, and market timing ,etc .Finding out the patterns and following them can be incredibly valuable for new startups.

Driving Informed Investment Decisions:

This can also help investors have better understanding on which industry or region to choose and make a better informed decision.

Supporting Startup Ecosystem:

Startup culture is booming day by day. This analysis can be of great use to the aspiring entrepreneurs to better understand the market demand and position themselves in way to attract fundings.

Data Modeling: Suitability for the Project

The dataset includes numerous categorical (e.g., industry type, geographic location) and continuous features (e.g., funding amounts). Based on these, Logistic regression approach would work.

WHY?

1) **Binary Classification for Startup Success**
2) **Interpretability of Influential Features**
3) **Probabilistic Output for Decision-Making**
4) **Efficient and Scalable for Large Datasets**
5) **Simple Implementation**

Features considered:

`Team Size`: A startup's team size often reflects its resource capacity and ability to scale operations, both of which can contribute to success. Larger teams may have more diverse skills, improving problem-solving and execution.

`Industry`: The industry type influences success since some sectors, like technology or finance, tend to have higher growth potential and investor interest. Analyzing industry patterns helps identify high-growth sectors

`Region`: The location or region is critical, as certain areas provide better access to funding, networks, and talent, which can enhance a startup's success chances.

`School`: The educational background, particularly prestigious institutions, can be associated with stronger networks and credibility, potentially attracting more investment and partnerships.

Features considered:

`Field of study`: Founders' areas of study can bring expertise relevant to their industry. For example, a tech-focused field of study is often advantageous for tech startups.

`Top company`: Having founders or team members from renowned companies adds credibility and experience, which investors often view positively, potentially boosting the startup's success.

Data Visualization

IDEA: Upcoming startups success rate prediction

Successful = educational background + founder prior company + penetration market region

If you come from a similar educational background or prior company with high success rate among startup companies, and not working with the very high precision of companies already in same geographical area but also not very low, you can penetrate the market more easily and likely succeed.

Data Visualization

Target Audience and Relevance to audience

- Investors who are looking for startups to invest
- New Entrepreneurs

Tools and Softwares

- Data Cleaning
  - Pandas
  - Numpy
- Data Analysis and Modeling
  - Sci-kit learn (GaussianNB, accuracy score and confusion matrix)
- Visualization
  - Matplotlib
  - Seaborn
  - Pyplot
  - Plotly

Data Visualization

Types of Visualization

- Choropleth maps or Cluster maps - geographical patterns
- Scatter Plots - education background likelihood of success
- Barcharts - to highlight importance of each features individually
- Donut chart

Performance Metrics for Visualization
How we will check the performance for the models?

- Confusion Matrix: how accurately model predicts, eg. false positives
- ROC Curve: predict 100 startups, how many actually succeed
- Precision-Recall Curve: how many success predicted accurately
- R-Squared value: how close your prediction is to actual value
- Mean Absolute Error: how far you are off from actual prediction

Data Visualization

Potential Issues with Visualization

Getting more technical,

- Cleanup data and Handle missing values (drop rows)
- Outliers - extreme values that can distort your results, use median
- Load all data in single data frame by joining with unique ids and hnids
- Convert categorical variables into numerical format (field_of_study, top_companies)
- Group fields like team_size to categories (ie. small, medium and large)
- Based on school and field of study, quantify educational background rating

Preliminary Results

Initial Data Characterization

The dataset has been categorised into multiple separate tables which is merged in single dataframe now for the project. The columns present are:

- Id
- Name
- Slug
- Website
- Small logo Url
- One Liner
- Longdescription
- Teamsize
- Url
- Batch
- Hnid
- Avatar_thumb
- Current_company
- Current_title
- Company_slug
- Top_company
- Company
- School
- Field_of_study
- Year
- Status
- Industry
- Region
- Country
- Address
- Tag
- Badge
- Firstname
- Last Name

Preliminary Results

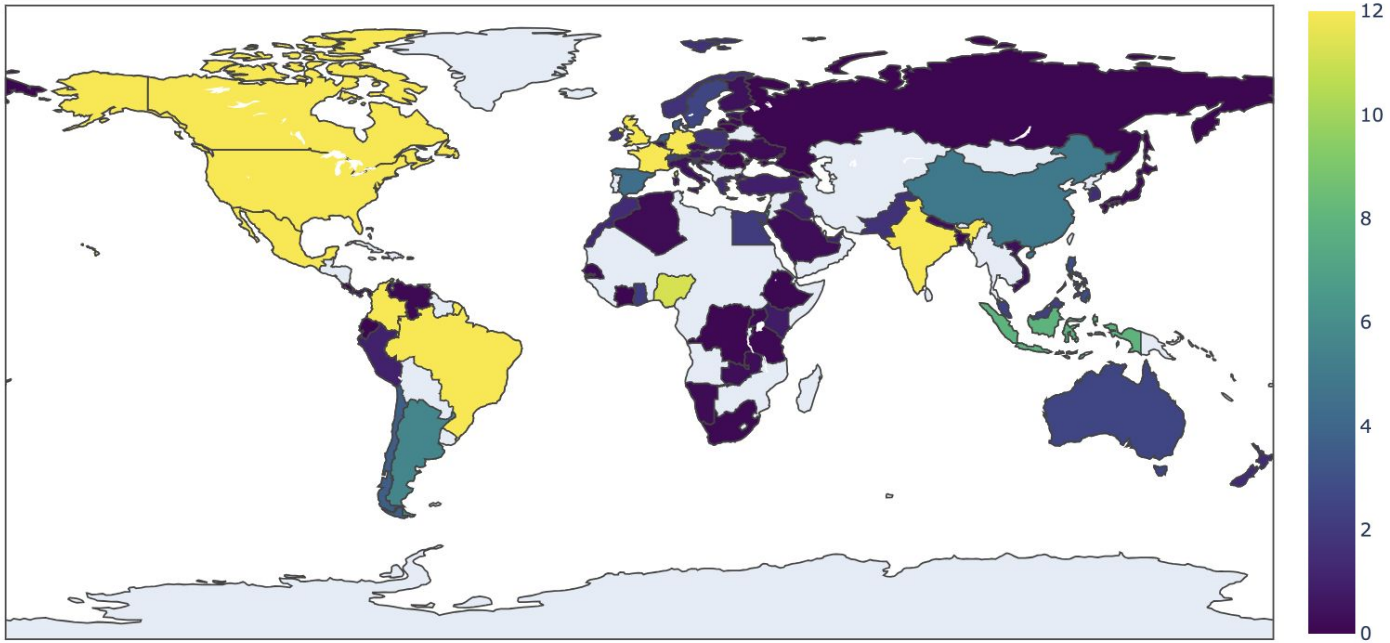
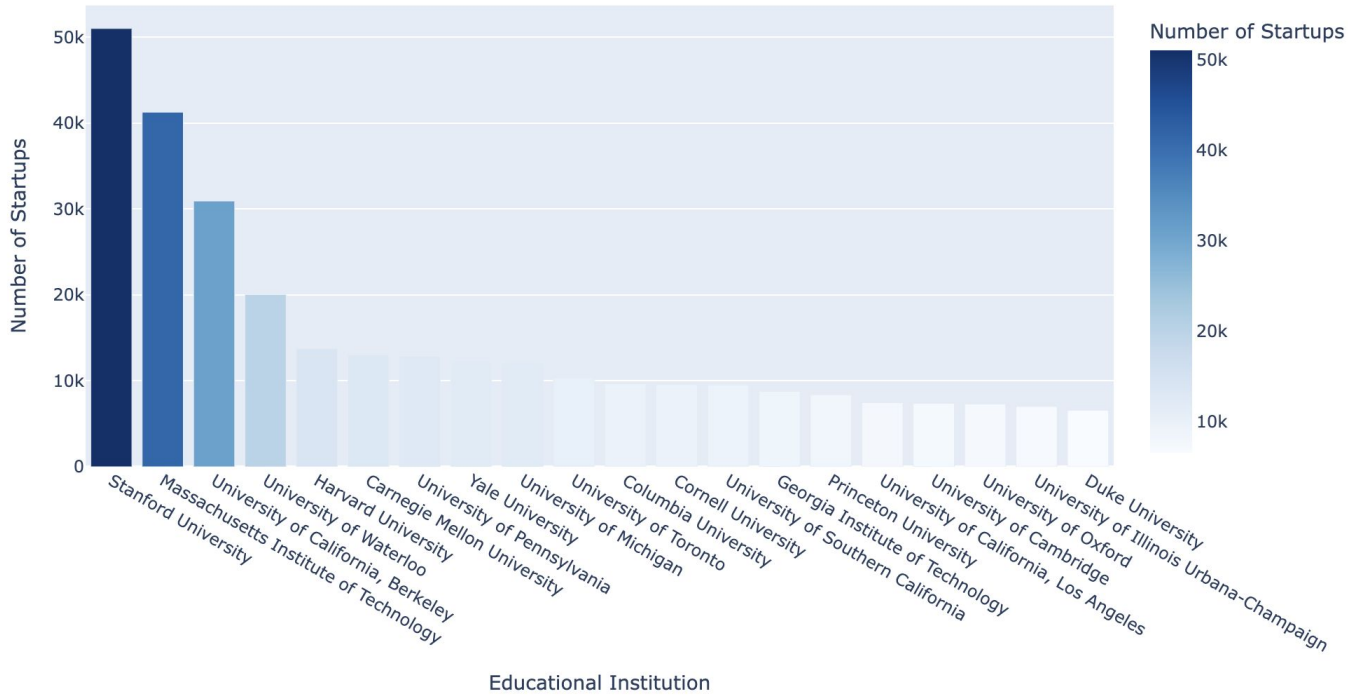


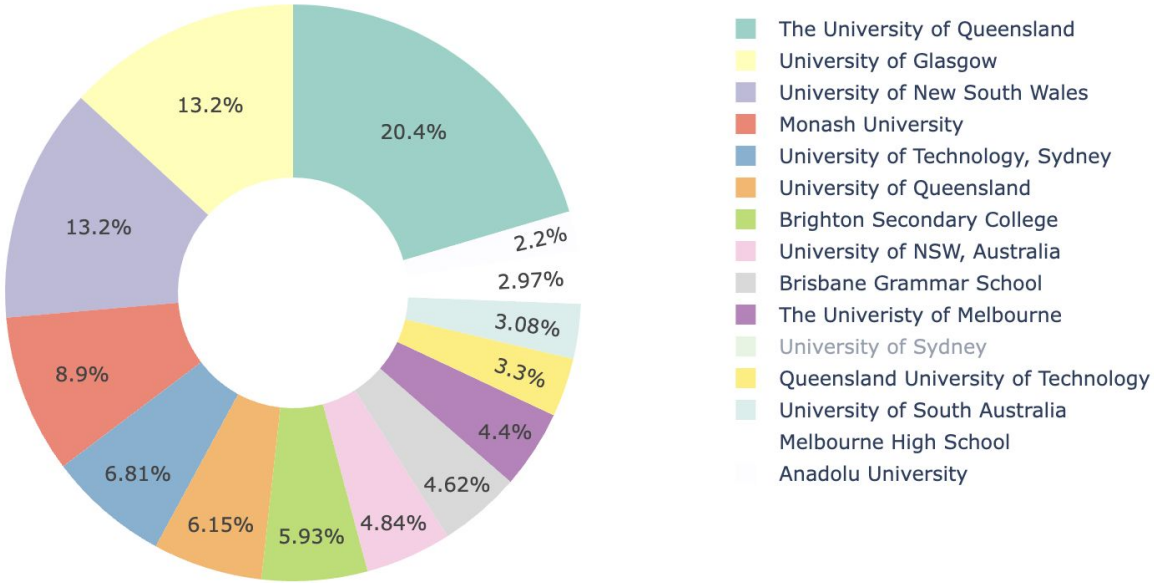
Fig: Startups from different countries

Preliminary Results



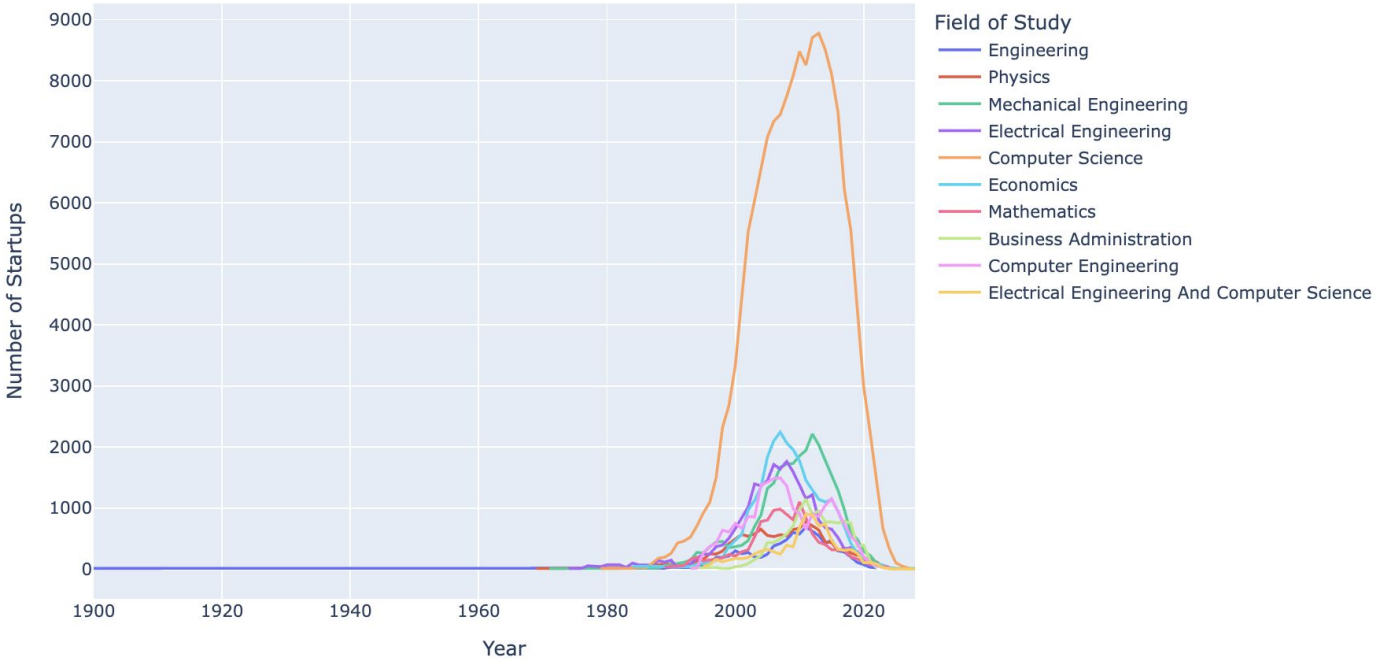
No of startups based on different universities

Preliminary Results



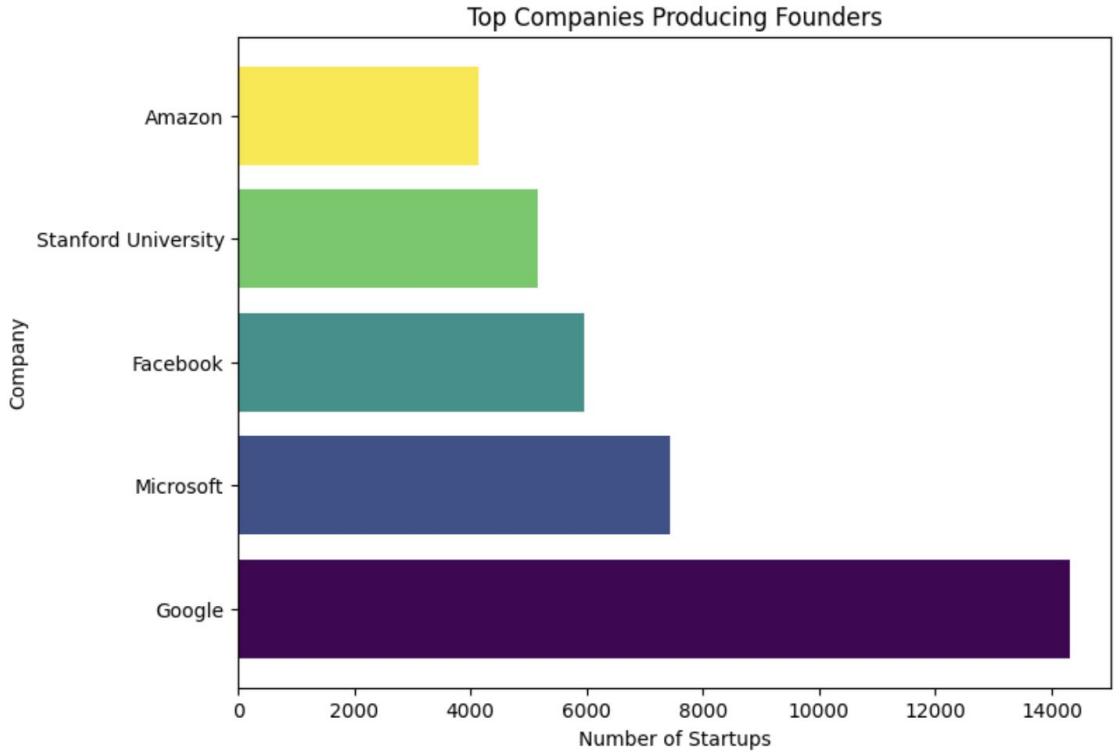
Top 15 universities with most number of founders

Preliminary Results

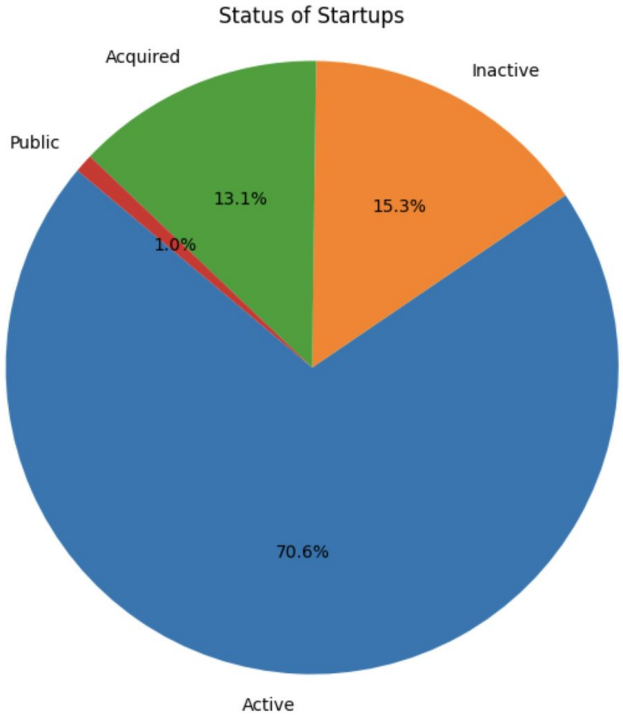


Founders field of study distribution by year graduated

Preliminary Results



Preliminary Results



Preliminary Results

Challenges Possible
Have you identified any limitations or areas for improvement based on the preliminary results?

- Imbalance data - oversampling or undersampling, status (inactive startups)
- Missing data
- Extreme values - team size
- Feature selection - which features to consider for more accurate prediction

Validation
What methods will you use to ensure the reliability and accuracy of your results moving forward?

- K-fold cross validation - Splitting the dataset into multiple parts and testing the model's performance on each part
- Train test split - Train test split, 80-20 split
- Confusion Matrix

References

Cohen, S., & Hochberg, Y. (2014). Accelerating startups: The seed accelerator phenomenon. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2418000

Marketeers. (2023, October 3). What is Y Combinator? - Marketeers. *Medium*. https://medium.com/@markeeters/what-is-y-combinator-1de598e83cd4

Shane, S., & Stuart, T. (2002). Organizational endowments and the performance of university startups. *Management Science*, *48*(1), 154–170. https://doi.org/10.1287/mnsc.48.1.154.14280

Sussan, F., & Acs, Z. J. (2017). The role of big data in startup ecosystems. *Small Business Economics*, *49*(2), 279–290. https://doi.org/10.1007/s11187-017-9874-9

Y Combinator Companies Directory. (n.d.). *Y Combinator*. Retrieved October 26, 2024, from https://www.ycombinator.com/companies

Time for Questions