New York University

Data Science for Business: Technical

Fall 2024: GB 2336 – Section 10

Final Report - Team 2

12/19/2024

Leon Jon

Srinidhi Jeyakumar

Trent Yu

Zitong Wang

Table of Contents

Enhancing VC Decisions:

Identifying High-Growth Startups with Machine Learning

1. Introduction

In the last decade, venture capital as an asset class has significantly outperformed public markets. As the ecosystem heads into a new decade, it's critical for investors to be proactive in their investment approaches. While investors often have robust ways to evaluate startups, a data-driven approach can help investors more effectively find great companies to invest in. Our project focuses on analyzing startup data to identify key features that can help us understand a company's likelihood of progressing to the growth stage. By comparing across companies and trying to identify key features that contribute to startup success can help investors more effectively "pick winners." Using machine learning techniques, our goal as data scientists is to help VC investors proactively understand and identify types of startups that have a higher potential for long-term success.

For this analysis, we used a dataset sourced from **Harmonic.ai**, a data platform specifically designed for venture capital investors. The dataset provides a rich collection of information on startups, including both **numerical features** (such as funding totals, headcount, and number of funding rounds) and **categorical features** (such as industry tags, customer type, and country). The dataset focuses on startups that have raised significant capital, covering those from **Series A to Series C+ funding stages**.

Our primary objective is to determine the most influential features that contribute to a startup reaching the **growth stage**. To achieve this, we used the CRISP-DM approach to identify a business problem, prepare the data and ultimately run our analysis.

2. Business Understanding

The venture capital (VC) industry faces a critical challenge in identifying startups likely to achieve significant funding and growth. This decision-making process is crucial for VCs, as it directly influences the allocation of limited resources, risk mitigation, and return on investment (ROI). However, traditional investment approaches rely on subjective judgment, leading to inefficient resource allocation and increased financial risk. By adopting a data-driven approach, VCs can improve the accuracy and efficiency of their decisions, focusing on startups with the highest potential and optimizing their investment strategies.

This project provides venture capital investors with a powerful tool for identifying and prioritizing high-potential investments. By leveraging predictive analytics, analysts can assess startups using measurable indicators, streamlining the evaluation process and reducing reliance on intuition. The model developed in this study highlights critical predictors of success, such as headcount growth and the number of funding rounds, enabling analysts to focus on evidence-based insights. This approach not only minimizes the risk of poor investments but also helps diversify portfolios for maximum ROI.

The value of this project lies in its ability to transform the investment decision-making process for VCs. It offers a scalable, adaptable, and efficient solution that aligns with the rapidly evolving demands of the startup investment landscape. By incorporating predictive analytics, the project supports VCs in making smarter, more informed decisions, ultimately contributing to the growth of promising startups and the success of the investment ecosystem.

3. Data Understanding

Data Source:

- **Source**: Harmonic.ai
- **Dataset Name**: Startup Data Across Series A Through C+
- **Description**:

  The dataset is provided by **Harmonic AI**, an actively managed data platform designed specifically for venture capital investors. Harmonic AI aggregates a combination of **publicly available** and **proprietary data** on startups, covering their lifecycle from stealth stage to growth stage (Series C+). It offers detailed information that supports investment research, including:

  - **Company Descriptions**
  - **Headcount**
  - **Founding Year**
  - **Industry Tags**
  - **Funding Information**
- **Scope**:

  While the full dataset contains over **27 million startups**, our project focuses specifically on companies that have raised significant capital and are at **Series A stage and beyond**. This subset is particularly relevant for identifying features that predict whether a startup can transition to the growth stage.

Dataset Overview

- **Total Entries**: 41,339 startups
- **Total Features**: 18 key features
- **Time Period**: Data includes startups across various years, focusing on those that have undergone Series A through C+ funding rounds.
- **Target Variable**:
  - **Stage**:
    - **Early Stage**: Startups in initial phases of growth.
    - **Growth Stage**: Startups that have demonstrated growth potential and have raised significant funding.

Feature Breakdown

Feature Name	Type	Description
Funding Total	Numerical	Total amount of funding received by the startup.
Headcount	Numerical	Number of employees in the startup.
Number of Funding Rounds	Numerical	The number of funding rounds the startup has gone through.
Funding per Employee	Numerical	Funding total divided by headcount, representing funding efficiency.
Last Funding Total	Numerical	Total funding amount received in the last funding round.

Country	Categorical	The country where the startup is based.
Customer Type	Categorical	The type of customer served (e.g., B2B, B2C).
Industry Tags	Categorical	Industry classifications (e.g., Biotechnology, Communications, Service).
Funding Stage	Categorical	Current funding stage (e.g., Early Stage, Growth Stage).
Last Funding Type	Categorical	The type of the most recent funding round (e.g., Series A, Series B).

Project Objectives

The main goal of this project is to **identify significant features that influence whether a startup reaches the growth stage**. By using machine learning models, we aim to:

1. **Determine which features** (e.g., headcount, funding amounts, industry) are most predictive of startup success.
2. **Provide actionable insights** for venture capital investors to prioritize startups with higher growth potential.

Initial Data Observations

1. **Distribution by Country**:
    - The majority of startups in the dataset are based in the **United States**.
    - Other notable countries include **China, the United Kingdom, India, and South Korea**.

2. **Funding Stage Distribution**:
    - **54.5%** of startups are in the **Early Stage**.
    - **45.5%** of startups are in the **Growth Stage**.
3. **Feature Correlations**:
    - **Funding Total** shows a positive correlation with **Headcount**.
    - **Industry Tags** like **Communications**, **Life Sciences**, and **Consumer Products** are associated with higher funding totals.

Data Challenges and Considerations

1. **Missing Values**:

    The dataset contained several missing values, particularly in categorical features like **Country** and **Industry Tags**. Different imputation strategies were applied based on feature types:

    - **Numerical Features**: Missing values in features like **Funding Total**, **Headcount**, and **Number of Funding Rounds** were replaced with the **median**.
    - **Categorical Features**: Missing values in features like **Customer Type**, **Industry Tags**, and **Country** were replaced with the label **'Unknown'** due to computational constraints.
2. **High-Null Columns**:
    - The **Technology Tags** feature contained a high percentage of missing values. Given its limited utility after imputation, it was dropped to maintain dataset integrity.
3. **Irrelevant Features**:

    - Unique or messy features such as **Last Funding Date**, **Company Name**, and **Company ID** introduced noise and complexity. These were dropped to simplify the analysis.

4. **Outliers**:
    - Outliers in **Funding Total** were identified and managed to prevent bias in the models.

4. Data Preparation

Step	**Description**	**Features Affected**	**Method Used**
Imputation of Missing Values	Replaced missing values with median (numerical) or mode (categorical).	Funding Total, Headcount, Country, Industry Tags	Median (numerical), Mode (categorical)
High-Null Column Removal	Dropped columns with excessive null values.	Technology Tags	Dropped
Irrelevant Feature Removal	Removed features with unique or messy values to reduce noise.	Last Funding Date, Company Name, Company ID	Dropped

Standardization	Standardized numerical features for consistency in modeling.	Headcount, Funding Total	Standard Scaling
Encoding Categorical Data	Converted categorical features to numerical values using one-hot encoding.	Country, Customer Type	One-Hot Encoding
Outlier Detection	Identified and handled outliers to avoid bias in the models.	Funding Total	Outlier Treatment

1. **Imputation of Missing Values**:
   - **Numerical Features**: For features like **Funding Total**, **Headcount**, and **Number of Funding Rounds**, missing values were replaced with the **median**.
   - **Categorical Features**: For features like **Customer Type**, **Industry Tags**, and **Country**, missing values were replaced with the label **'Unknown'** due to computational limitations for more advanced imputation methods like K-Nearest Neighbor (KNN).
2. **Dropped High-Null Columns**:
   - The **Technology Tags** feature was removed because it contained too many missing values and would not contribute meaningful insights.
3. **Dropped Irrelevant Features**:
   - Columns like **Last Funding Date**, **Company Name**, and **Company ID** were dropped as they introduced noise and were difficult to group.

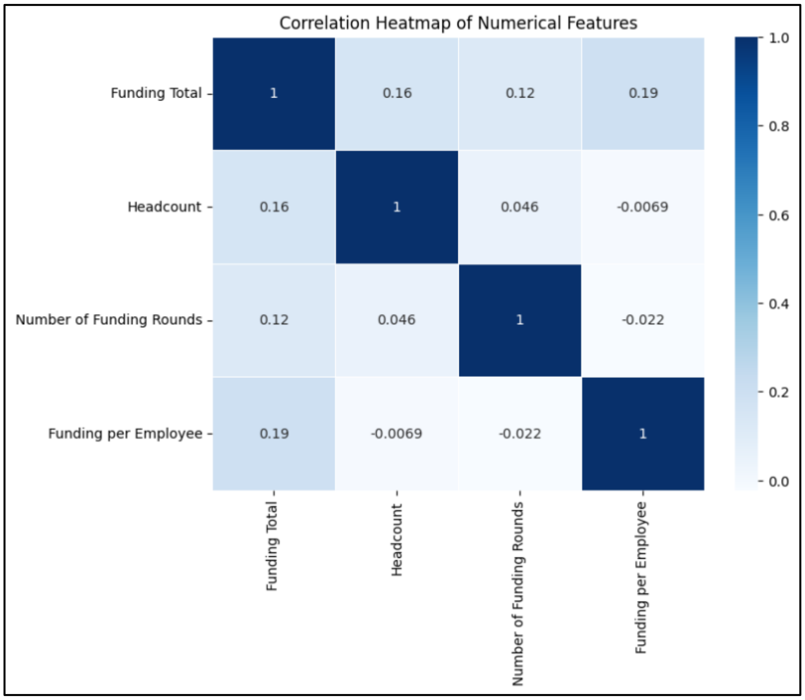
4. **Standardization**:
    - Numerical features were standardized using **Standard Scaling** to ensure consistency.
5. **Encoding**:
    - **One-hot encoding** was used for categorical variables like **Country** and **Customer Type**.
6. **Outlier Detection**:
    - Managed outliers in **Funding Total** to prevent them from disproportionately influencing the machine learning models.

After the data cleaning process, the final dataset consisted of:

- **Total Entries**: 41,339
- **Features**: 15 (a mix of standardized numerical features and one-hot encoded categorical features)

5. Exploratory Data Analysis

5.1 Correlation heatmap of key numerical features.

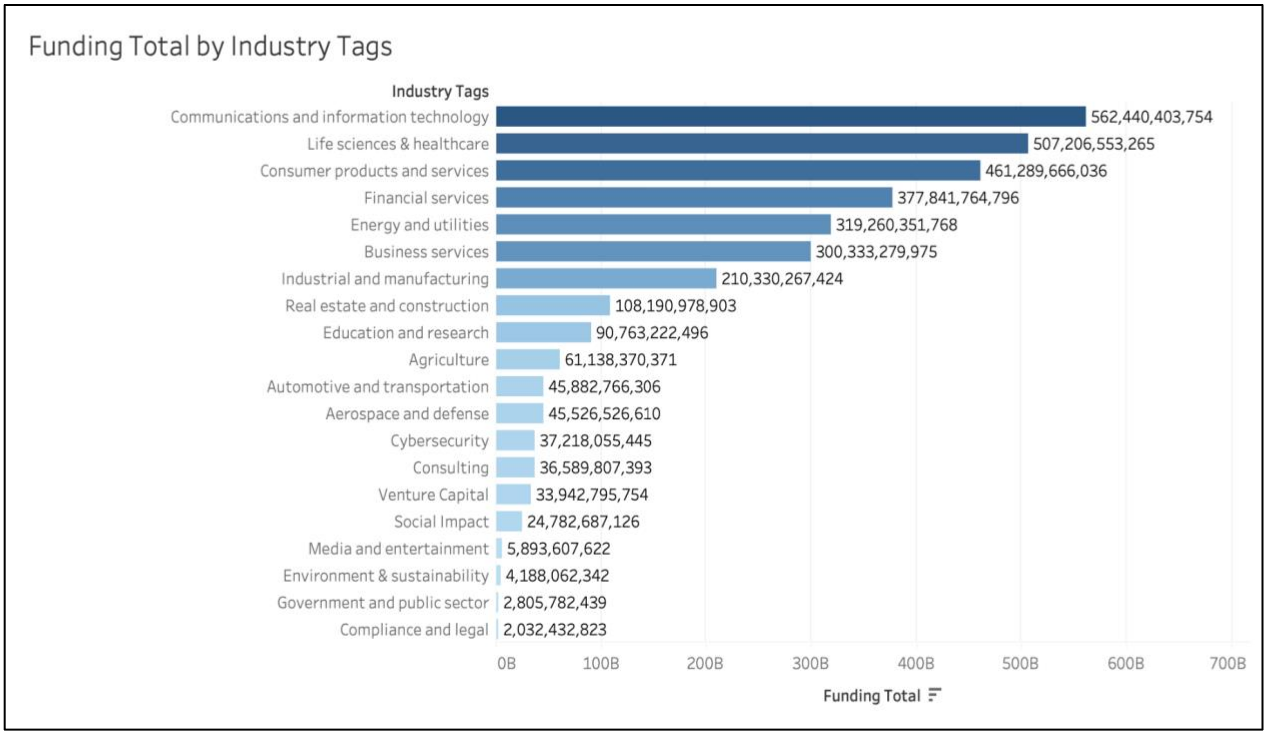


Insights:

- **Funding Total** has a weak positive correlation with **Headcount** (0.16), **Number of Funding Rounds** (0.12), and **Funding per Employee** (0.19).
- **Headcount** shows very little correlation with the other variables, indicating that headcount alone may not be a strong predictor of total funding.
- **Number of Funding Rounds** has a minimal relationship with the other variables.

This suggests that while these features provide some predictive value, they are not highly correlated, which could be beneficial for machine learning models to avoid multicollinearity issues.

*5.2 **Bar chart showing funding totals across different industry tags.***

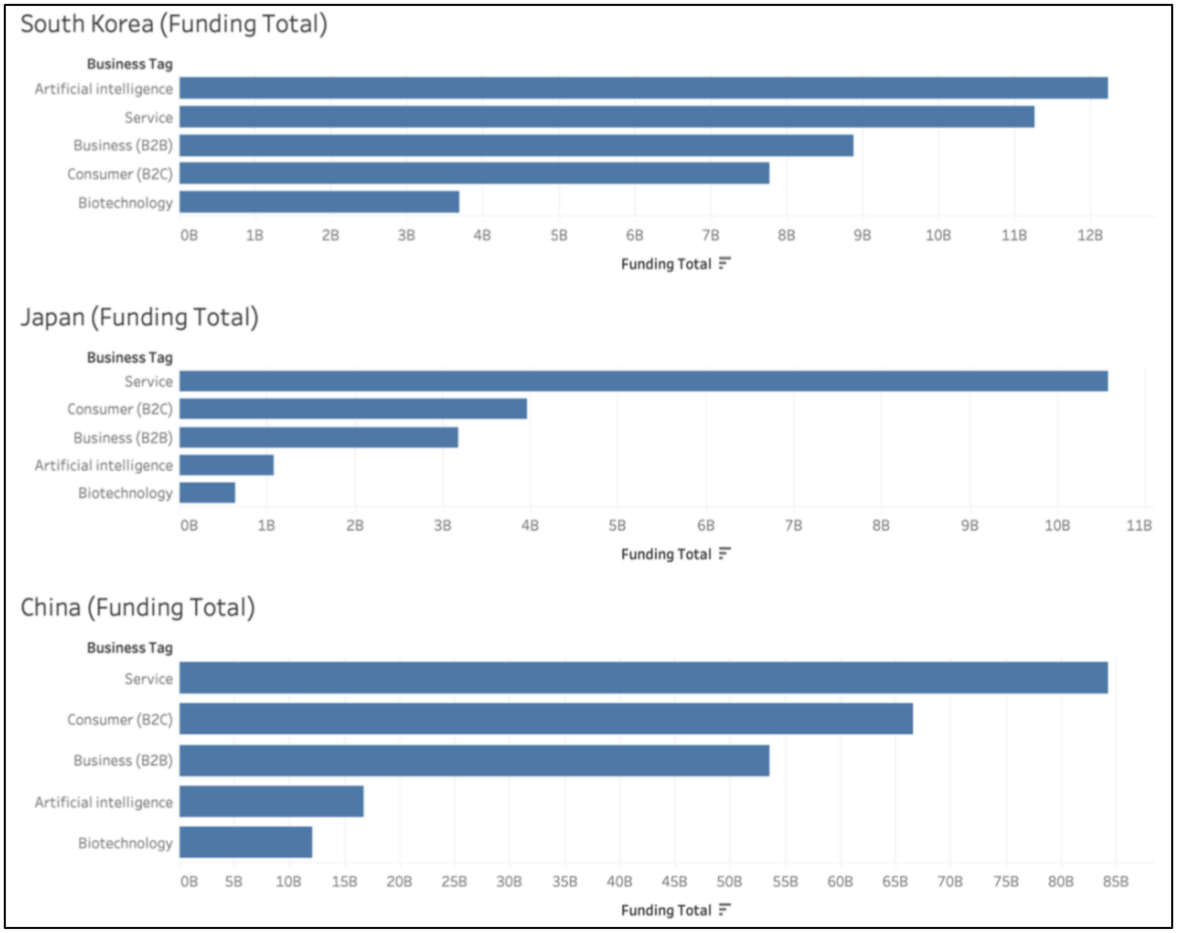


Insights:

- The **Communications and Information Technology** sector received the highest funding, totaling over **$562 billion**.
- Other high-funding industries include **Life Sciences & Healthcare** ($507 billion), **Consumer Products & Services** ($461 billion), and **Financial Services** ($377 billion).
- Industries such as **Energy & Utilities** and **Business Services** also attracted significant investments.
- In contrast, industries like **Environment & Sustainability** and **Government & Public Sector** received comparatively lower funding.

This analysis highlights which industries are more likely to secure significant funding, helping to identify potential areas of growth and investment opportunities.

5.3 Bar charts displaying funding totals for different business tags in South Korea, Japan, and China.

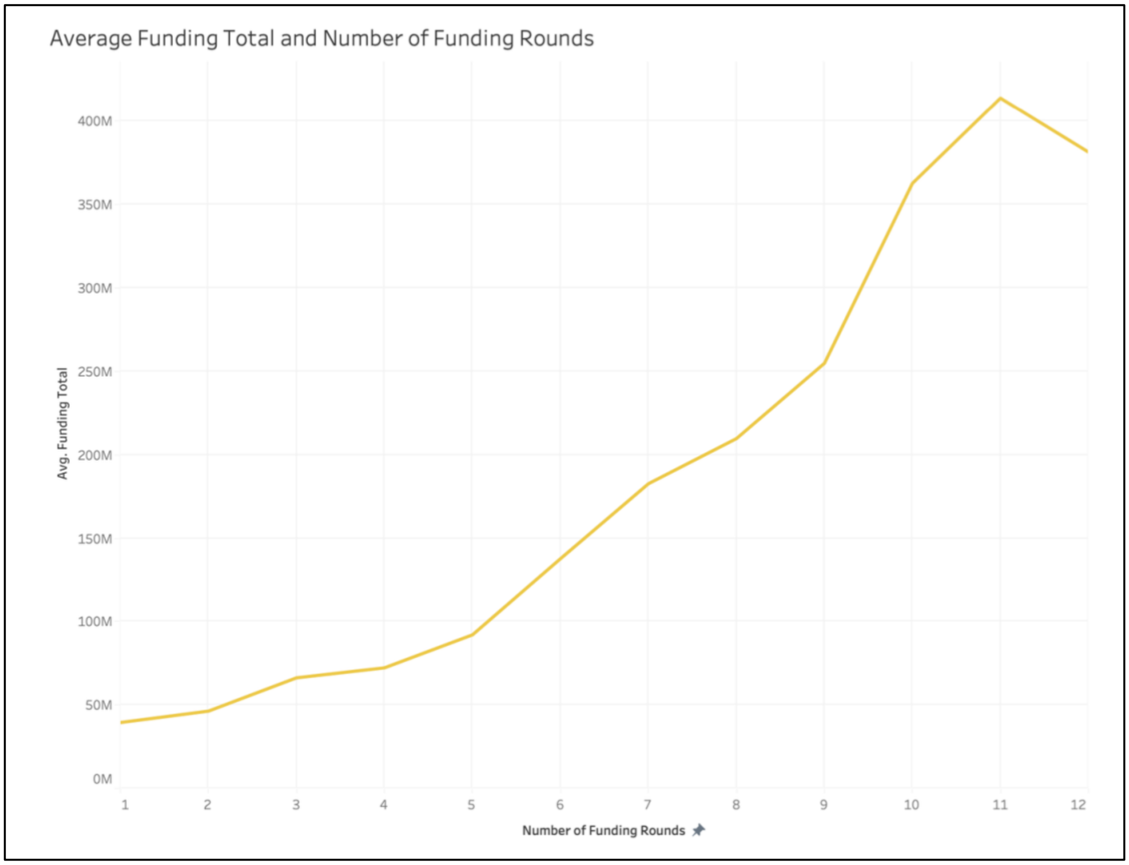


Insights:

- In **South Korea**, the **Artificial Intelligence** and **Service** sectors received the highest funding, each exceeding **$12 billion**.
- In **Japan**, the **Service** sector leads with funding totals exceeding **$10 billion**, followed by **Consumer (B2C)** and **Business (B2B)** sectors.
- In **China**, the **Service** sector dominates, with over **$80 billion** in funding. Other well-funded sectors include **Consumer (B2C)** and **Business (B2B)**.

These insights suggest that different countries have distinct investment focuses, which can help tailor investment strategies and identify regional opportunities for startups.

5.4: Line chart showing the average funding total as the number of funding rounds increases.



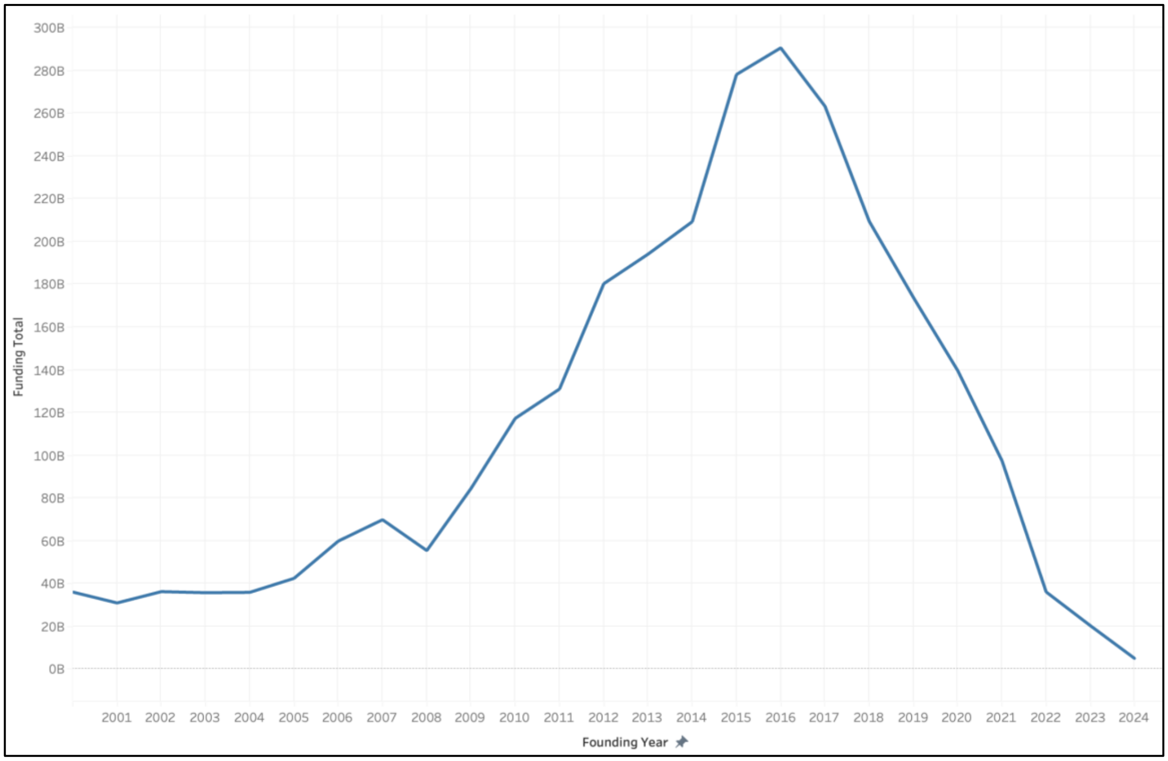
Insights:

- There is a **positive correlation** between the number of funding rounds and the average funding total.
- Startups with **more funding rounds** tend to receive significantly higher total funding.
- The average funding increases steadily from around **$40 million** for startups with one funding round to over **$400 million** for those with 11 or more funding rounds.

- A sharp increase is observed after the **9th funding round**, suggesting that startups reaching this stage have a higher likelihood of securing substantial investments.

This trend underscores the importance of securing multiple funding rounds, indicating that companies with a track record of successful funding rounds are more likely to progress to the **growth stage**.

5.5: Line chart showing the total funding by founding year.



Insights:

- There was a steady increase in funding totals from **2005 to 2015**, with a peak around **2015 to 2017**.
- After **2017**, funding totals began to decline sharply, with a significant drop-off observed between **2019 and 2022**.

- This trend may reflect broader market dynamics, such as economic cycles, shifts in investor sentiment, or changes in startup activity levels.

Understanding these trends helps contextualize the funding landscape and identify periods of high growth or contraction.

6. Modeling and Evaluation

With the dataset cleaned, preprocessed, and ready for analysis, we now turn our attention to the machine learning approach. After addressing missing values, removing high-null and irrelevant features, and standardizing the data, we have a dataset that balances both **numerical** and **categorical features** effectively. The next step involves applying appropriate machine learning algorithms to predict which startups are likely to reach the **Growth Stage**.

Since the dataset consists of both numerical and categorical variables, we decided to use machine learning algorithms that can natively handle both types of variables together. We ultimately settled on the following algorithms for our analysis:

- **Random Forest:** An ensemble algorithm that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. It can natively handle both categorical and numerical variables, making it versatile for datasets like ours that include mixed data types.
- **Decision Tree:** Splits data into branches based on feature values, handling both categorical and numerical variables natively without requiring transformation. It is interpretable and practical for understanding relationships in datasets with mixed data types.

- **Logistic Regression:** Often used for binary or multi-class classification by estimating the probability of an outcome using the logistic function.
- **Gradient Boosting:** An iterative ensemble method that builds decision trees sequentially, where each new tree corrects errors made by previous trees by minimizing a loss function.
- **XGBoost:** An optimized and efficient implementation of Gradient Boosting designed for high performance and speed. It uses advanced regularization techniques to prevent overfitting.

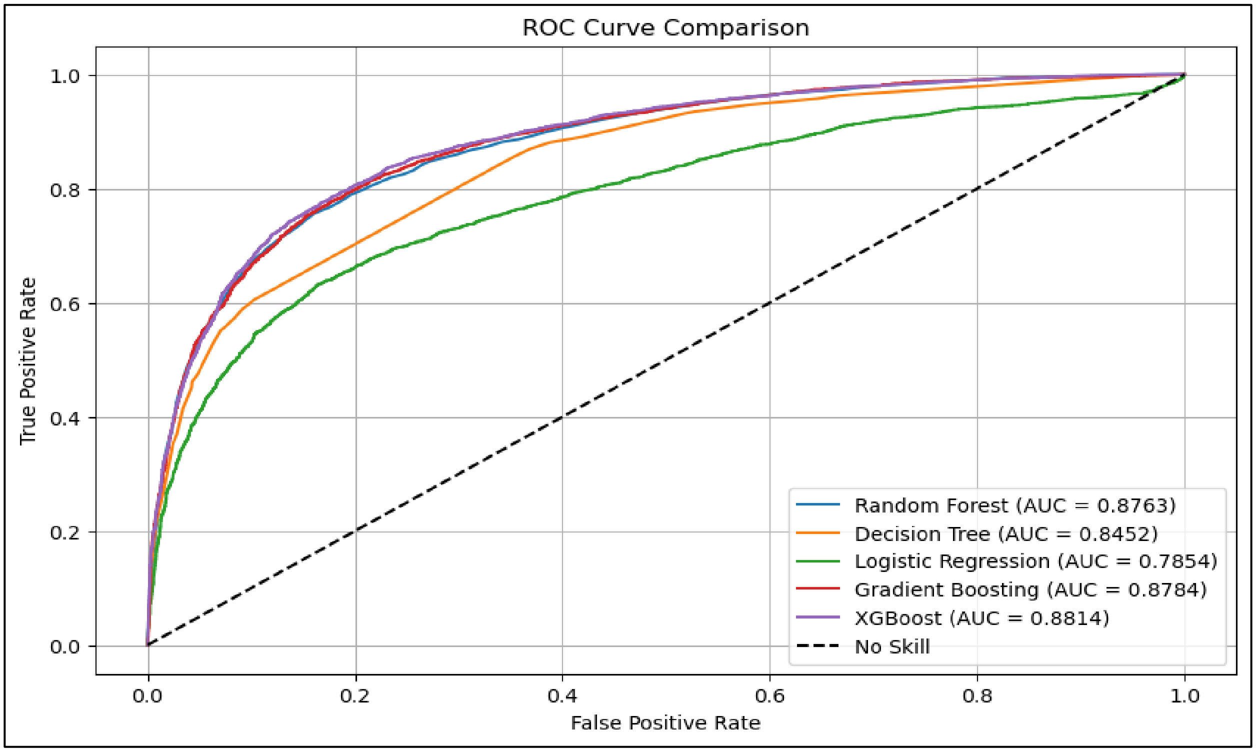
Target Variable:

- First of all, we set the feature, “Stage” as the target variable. The feature contains several stages based on the funding round and we believe that companies at the higher stages are more likely to grow up next term with additional funding. The scale of the stage starts from Stage A to Stage H and we divided the feature into two groups based on each company’s stage. Every company belonging to Stage A is marked as ‘Early Stage’ and the others are marked as ‘Growth Stage’. Since we targeted companies at the ‘Growth Stage’, target = ‘Growth Stage’ became our target variable.

Model comparison using AUC (Area Under the Curve):

- We wanted to compare the models’ performance that we chose using AUC. We built a classifier using an 80/20 split. After the first trial, we found that AUC for the models were nearly 1 and it seemed suspicious. We calculated the feature importance and found out that one of the features, ‘Last Funding Type’, has a significantly high feature importance. It might have caused data leakage. This can happen when the feature inadvertently includes information about the target variable, so it can artificially inflate performance. After dropping the feature, the performance of the model became reasonable.

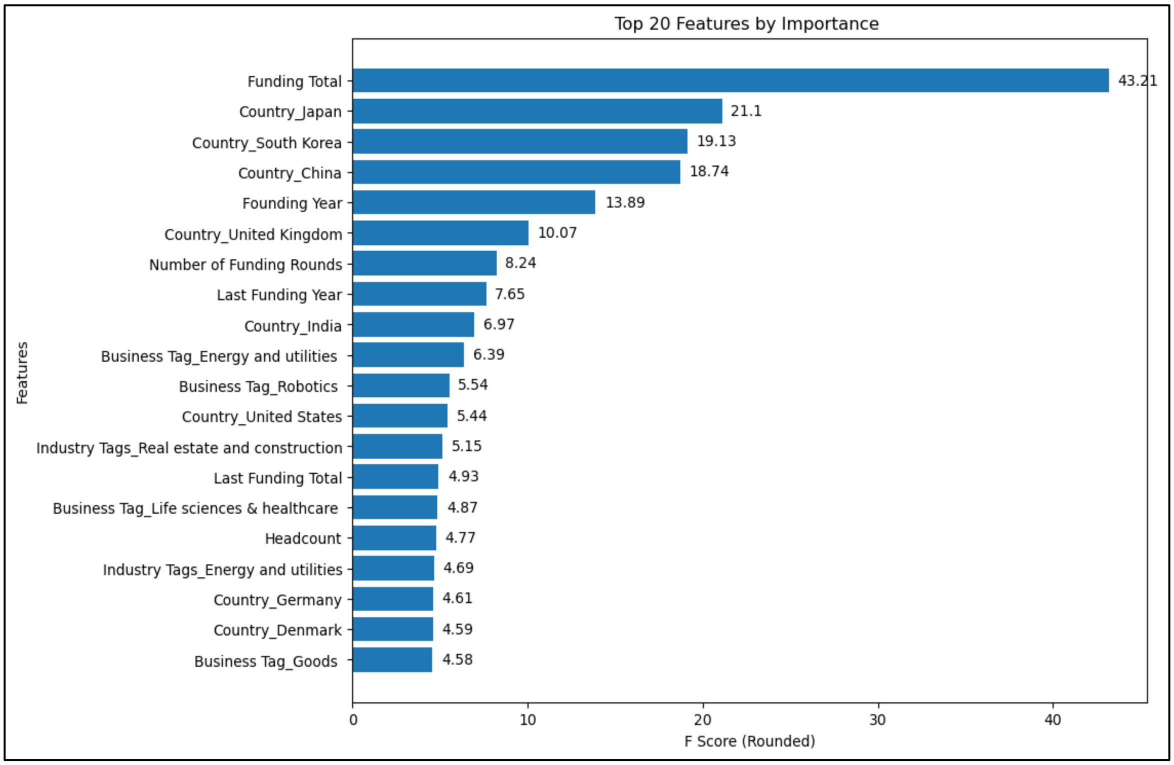
- The following figure shows the ROC Curve Comparison with the models we selected.



According to the model comparison, XGBoost performs better than other models. Random Forest, and Gradient Boosting also showed high performance but we decided to stick with XGBoost for the rest of the analysis.

Feature Importance:

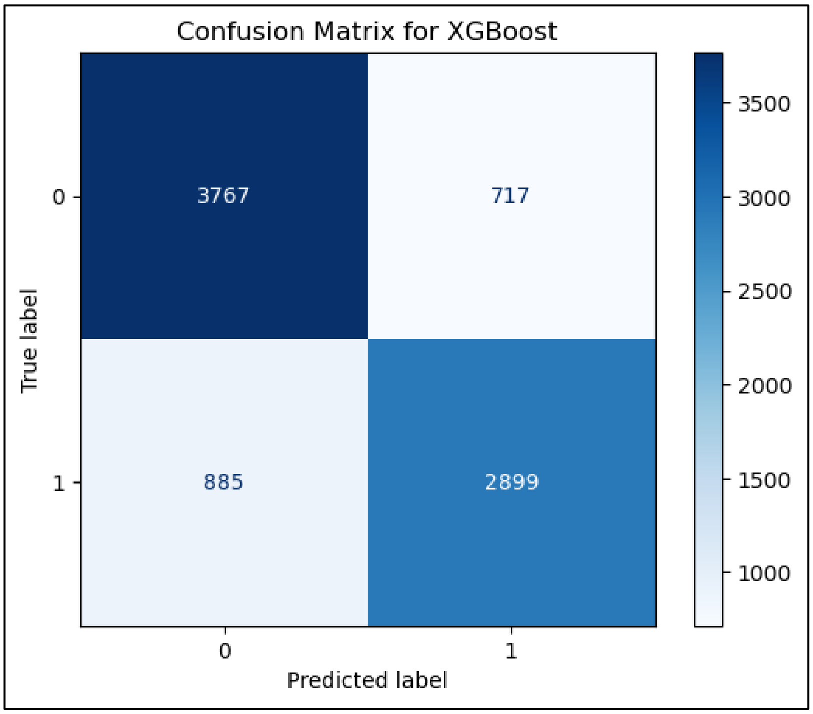
After selecting XGBoost as the main ML algorithm for the rest of the analysis, we calculated the feature importance to find out which features are more important for startup companies to get into the growth stage level. The following feature shows the top 20 features by importance.



The feature shows that ‘Funding Total is the most important feature for startup companies to get into the ‘Growth Stage’ level. Following the lead feature, Japan, South Korea, China turned out to be one of the most important features. We assume that startup companies from East Asian countries are fully supported by their governments and have a great environment for startup companies to grow up.

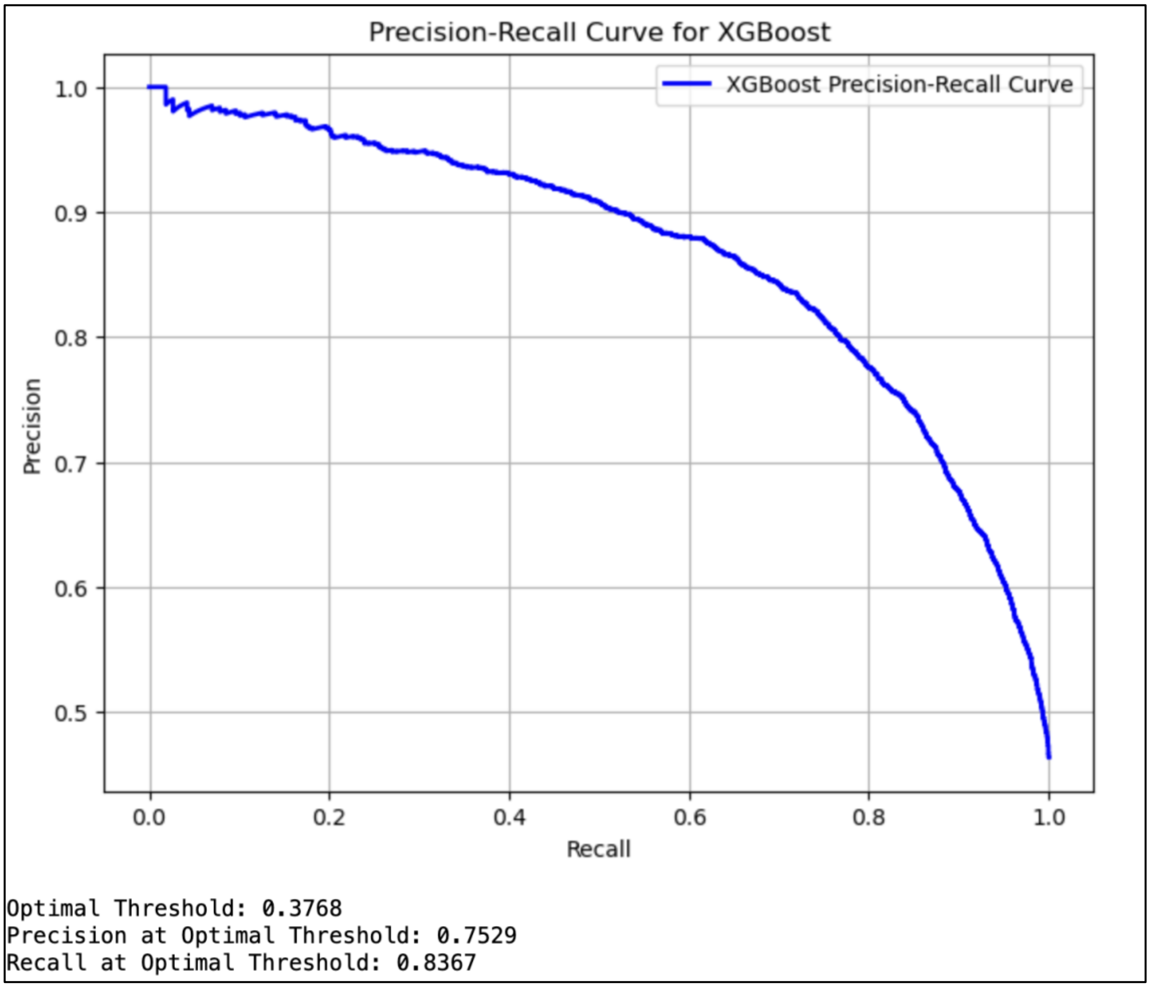
Confusion Matrix:

Based on the analysis, we then created a confusion matrix:



- ***True Negative (TN = 3767):***
  - Companies correctly predicted not to be in the "Growth Stage."
- ***False Positive (FP = 717):***
  - Companies incorrectly predicted as being in the "Growth Stage" but are not legit potential startups and this is what we want to avoid.
- ***False Negative (FN = 885):***
  - Companies that are actually in the "Growth Stage" but were predicted as not in Growth Stage.
- ***True Positive (TP = 2899):***
  - Companies are correctly predicted to be in the "Growth Stage."

As a VC, we should prioritize avoiding False Positives (FB) because an FB means we might mistakenly invest in a company that doesn't have legitimate growth potential. Also, this can lead to financial losses or poor investment returns. The 717 False Positives in the matrix represent cases where companies are predicted as "Growth Stage" but are not in the Growth Stage. To find out the optimal threshold, we made a Precision-Recall Curve for XGBoost and the following feature shows the result.



```
Optimal Threshold: 0.3768
Precision at Optimal Threshold: 0.7529
Recall at Optimal Threshold: 0.8367
```

Since we should focus more on avoiding False Positives, we want to increase the threshold from 0.3768 to 0.6. As a result, we got 0.84 in precision and 0.7 in recall. Increasing the threshold will reduce False Positives but might miss some true “Growth Stage” companies as it increases False Negatives.

7. Impact of Work

The impact of this project is significant, as it provides venture capital firms with a data-driven approach to streamline their investment decisions and improve outcomes. By identifying startups with the highest growth potential, the predictive model can help VCs allocate resources more effectively, reducing financial risks and increasing returns on investment. For example, by prioritizing startups with strong indicators such as headcount growth and multiple funding rounds, VCs can target companies statistically more likely to succeed, potentially increasing portfolio profitability.

Quantifying this impact, we estimate that reducing false-positive investments—where resources are allocated to startups with low growth potential—could save millions of dollars annually, depending on the scale of the firm's operations. For instance, if the model reduces false positives by 20%, a firm investing $100 million annually could potentially redirect $20 million to more promising ventures, leading to higher returns. Additionally, increasing the precision of predictions enhances the firm’s ability to build diversified, high-performing portfolios, which could yield long-term financial growth.

Beyond financial benefits, the project’s impact extends to operational efficiency. Automating the initial evaluation process saves time and reduces labor costs for analysts, enabling them to focus on higher-level strategic decisions. Furthermore, the insights provided by the model could foster stronger relationships with investors and stakeholders by demonstrating a commitment to data-backed decision-making, reinforcing trust and credibility within the investment community. This combination of financial, operational, and reputational advantages highlights the transformative potential of this work for the venture capital sector.

8. Deployment and Future Work

The deployment of the data mining results would involve integrating the predictive model into venture capital firms' decision-making processes. This can be achieved by building a user-friendly interface, such as a dashboard, where analysts can input key indicators about a startup and receive predictions about its likelihood of reaching the growth stage. This tool would provide real-time insights, allowing decision-makers to quickly evaluate potential investments and prioritize those with the highest predicted success.

After deployment, the model must be actively monitored to ensure its ongoing accuracy and relevance. Monitoring could include tracking the model's performance metrics, such as precision, recall, and overall accuracy, by comparing predictions with actual outcomes over time. Regular retraining of the model would also be necessary as new data becomes available, ensuring it adapts to market conditions and startup dynamics changes. Additionally, a feedback loop could be established where analysts provide input on the model's recommendations, helping to refine and improve its functionality.

The full Harmonic dataset contained over 27 million unique (companies). Given our computational constraints, we didn't include the much earlier-stage companies (e.g., stealth, pre-seed, and seed stage companies) that have raised little or no funding. Given more time and resources, we would have spent more time preparing and enriching the dataset to better understand company performance across the entire lifecycle of a startup. It would also be interesting to access the investor database to learn about the “best” investors (i.e., the ones who invest in companies that end up exiting

Additional datasets could provide valuable insights and enhance the model’s performance for future analyses. For example, sector-specific data, such as industry growth rates, market trends, or competitive landscape information, could help tailor predictions to particular industries. Social media engagement metrics, patent filings, or detailed financial data beyond funding totals could offer additional insights into a startup's defensibility. Expanding the dataset to include international market trends and economic indicators could help capture the global dynamics influencing startup success, creating a more comprehensive tool for venture capital decision-making.

9. Conclusion

In this project, we used machine learning approaches to help investors identify features of startups with high growth potential, using a dataset sourced from Harmonic.ai.

Through our analysis, we identified key factors influencing startup growth, such as **total funding**, **number of funding rounds**, and a greater concentration of startup activity in **geographic regions** like Japan, South Korea, and China. The **XGBoost model** emerged as the best-performing model, providing reliable predictions with an **AUC score of 0.8814**. We adjusted the decision threshold in order to balance precision and recall, minimizing false positives, and improved the mode’s performance to ensure better investment decisions.

The learnings from this project sets the foundation for further refinement and data inclusion that would yield richer insights for investors. We're excited to continue building on our existing work and explore future deployments.

Appendix

- **Code: Google Colab**
- **Data file: Google Drive**
- **File name: Dataset for ML.csv**

Team Member	**Contribution Area**	**Details of Contributions**
Leon Jon	**Data Understanding**	- Researched and documented the dataset from Harmonic.ai.
	Data Preparation	- Handled missing value imputation (numerical & categorical).
		- Dropped high-null columns (*Technology Tag*).
		- Removed irrelevant features (*Company Name*, *Company ID*, *Last Funding Date*).
		- Standardized numerical data and applied one-hot encoding to categorical data.
	Analysis	- Participated equally in data analysis.

Srinidhi Jai	**Exploratory Data Analysis**	- Created visualizations (correlation heatmap, bar charts, line charts).
		- Identified patterns in funding, geography, and industry sectors.
	Report Writing	- Wrote the **Exploratory Data Analysis, Data Understanding and Data Preparation** section.
	Analysis	- Participated equally in data analysis.
Trent Yu	**Modeling and Evaluation**	- Selected and implemented ML models (Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, XGBoost).
		- Addressed data leakage, evaluated models using AUC, and identified XGBoost as the best model.
		- Created ROC curve, confusion matrix, and precision-recall curve.
	Report Writing	- Wrote the **Modeling and Evaluation** section.
	Analysis	- Participated equally in data analysis.
Zitong Wang	**Report Compilation**	- Integrated sections into the final report and ensured compliance with guidelines.
	Presentation	- Created presentation slides and coordinated rehearsals.
	Analysis	- Participated equally in data analysis.