Startup Success Prediction Using Machine Learning

name: Qiyuan Zhu, Zella Yu

1. Problem to Tackle

Our project focuses on predicting whether a startup will remain active (including acquired and IPO) or eventually close. We plan to use company data such as founding year, total funding, location, number of funding rounds, and investment timing to estimate the likelihood of survival.

Startup outcomes are influenced by many factors like funding environment, team background, and market timing, which makes success prediction highly uncertain. Our goal is to use data-driven methods to explore these patterns and help investors and founders make more informed decisions.

---

2. Why This Is an Interesting and Useful Application

We find this topic interesting because predicting a company's success or failure has always been a challenging problem. It's often hard for people to judge outcomes based only on experience, since success depends on many factors that are difficult to evaluate together.

Using data science allows us to process large-scale data and identify relationships that might not be obvious to humans. Machine learning models can help uncover patterns in funding, timing, and location that relate to startup survival. Overall, data science provides an objective, systematic way to study what makes startups succeed or fail.

---

3. Datasets & Models

We start with the **Y Combinator Companies Dataset (2005–2024)**, focusing on 1,466 companies that already have clear outcomes. We label *Acquired/Public* as success and *Inactive* as failure.

Key features include founding year (converted to company age), industry type, geographic region, YC batch, and team size. We handle missing values by filling them with the median or "unknown," merge rare categories, and use one-hot encoding for categorical data.

Later, we plan to add the **Crunchbase Startup Success/Fail Dataset (~66K firms)**, which includes richer information such as total funding amount, investor count, and founder profiles.

Planned Models:

- Logistic Regression – interpretable baseline model.
- Random Forest – captures non-linear relationships.

- XGBoost / LightGBM – handles complex feature interactions through boosting.
 After expanding the dataset, we also plan to test simple ensemble and tuning methods to improve model accuracy.

---

4. Preliminary Performance

We tested three basic models on the YC dataset using an 80/20 train-test split. All of them performed better than a random baseline (about 58% accuracy).
 Random Forest achieved the best overall balance between precision and recall, while XGBoost gave slightly higher precision. These early results suggest that even simple models can learn useful patterns for predicting startup outcomes.

Model	Accuracy	F1 (Success)
Logistic Regression	63.9%	0.47
Random Forest	**67.3%**	**0.57**
XGBoost	67.3%	0.55

---

5. User Interface Design

We plan to build a simple **Streamlit dashboard** so users can explore model results and make their own predictions.

- **Visualization:** dropdown menus to switch between charts showing performance and feature importance.
- **Regional Analysis:** option to view predictions for different regions (e.g., U.S. vs. Europe).
- **Interactive Prediction:** users can enter company info such as industry, region, and funding to get an estimated survival probability.
 We hope this interface will make the model more engaging and easier to understand for non-technical users.