Startup Status Classification

by Startup Lens

Problem Statement: Many startups fail within their first few years, creating a need to predict whether a startup will remain **Active** or become **Inactive**. ***This prediction relies on factors like financial performance, growth metrics, funding, and market trends.***

Why is this Significant?

- **High Failure Rate**: A significant percentage of startups fail, resulting in substantial financial losses and wasted resources.
- **Investment Efficiency**: Early prediction helps investors make more informed decisions, optimizing capital allocation.
- **Data-Driven Insights**: Providing startups with actionable insights helps improve their chances of success.

Impact of Solving This Problem:

- **Reduced Investment Risk**: Investors can make better decisions, minimizing losses.
- **Higher Startup Survival**: Startups can adjust strategies early, leading to increased survival rates.
- **Economic Growth**: Successful startups contribute to innovation and job creation, fostering economic growth.

Dataset Overview

Dataset name: StartUp Investments (Crunchbase)

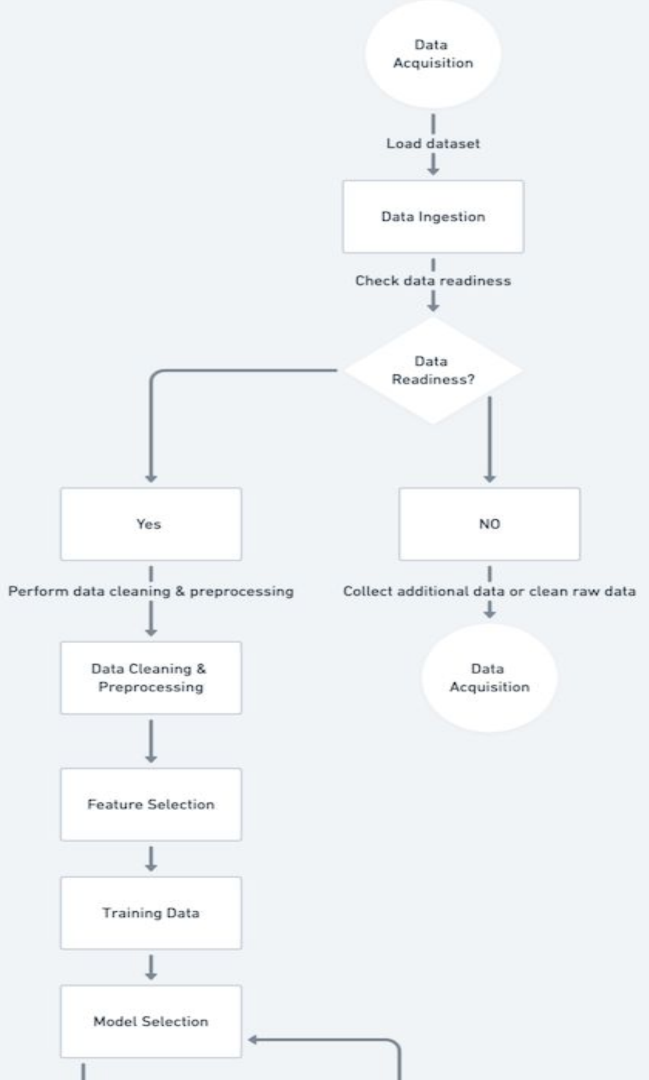
Dataset link: https://www.kaggle.com/datasets/arindam235/startup-investments-crunchbase

Dataset size: *54294 rows and 39 columns*

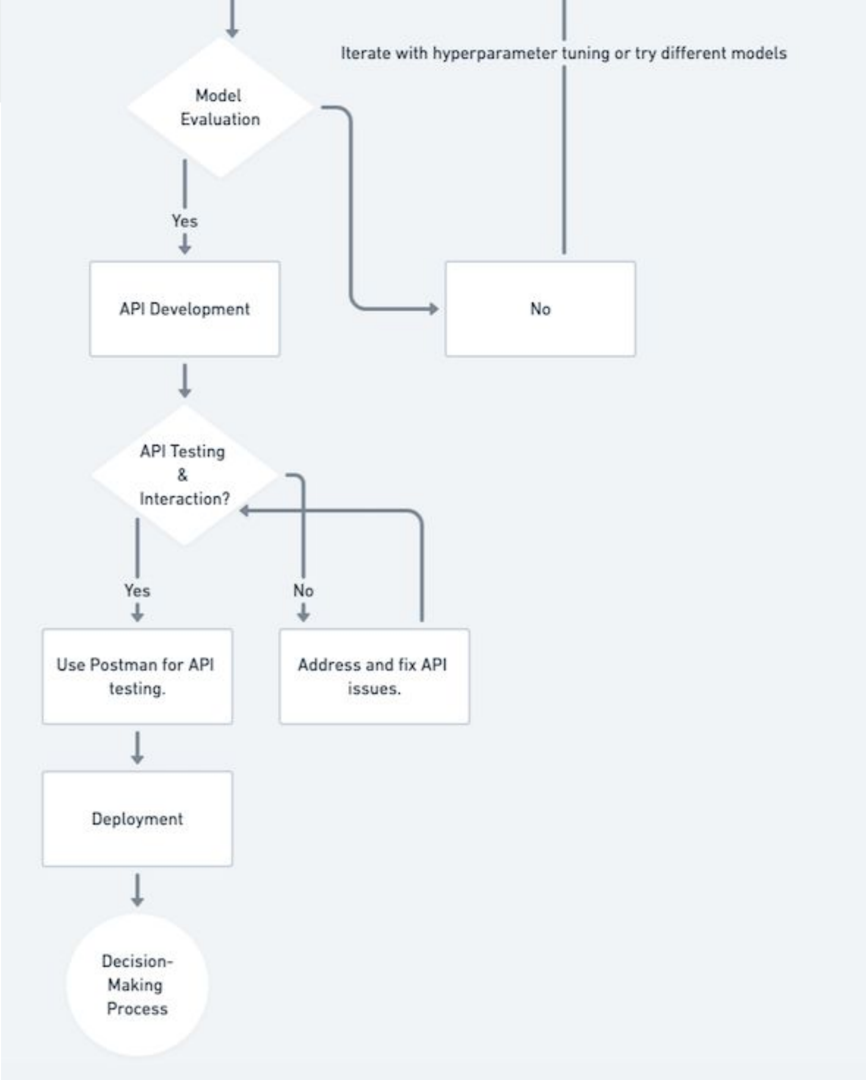
Why we chose this dataset ?

As startups are growing rapidly these days. We wanted to make a model which will predict the success rate of the startup. This dataset is useful because it provides real-world information on factors that drive startup success or failure, such as financials, funding rounds, and growth metrics.

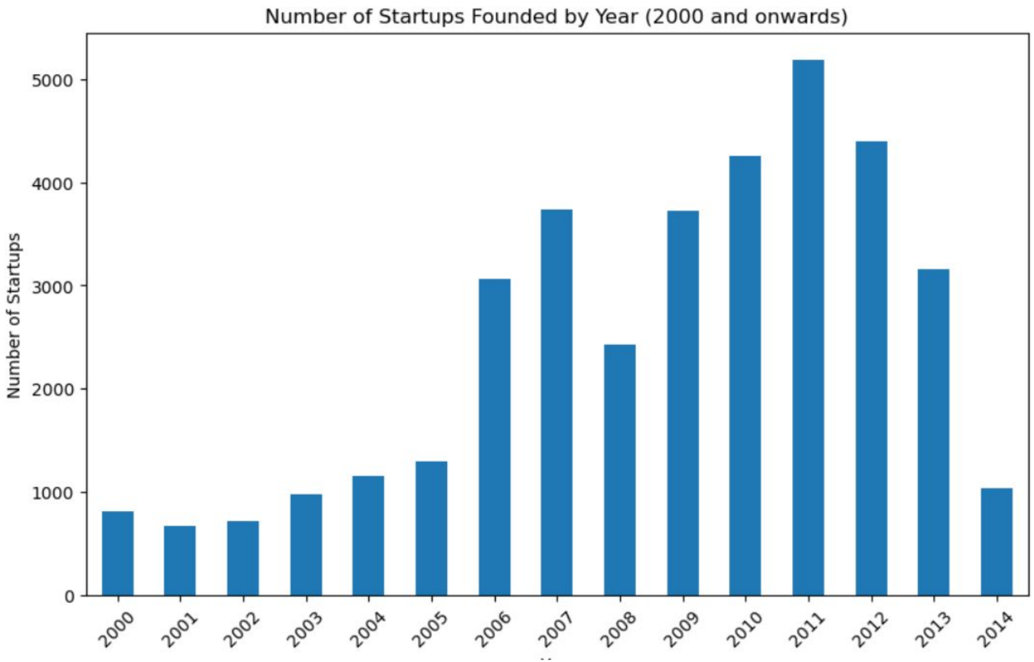
Flow Chart Diagram



Flow Chart Diagram



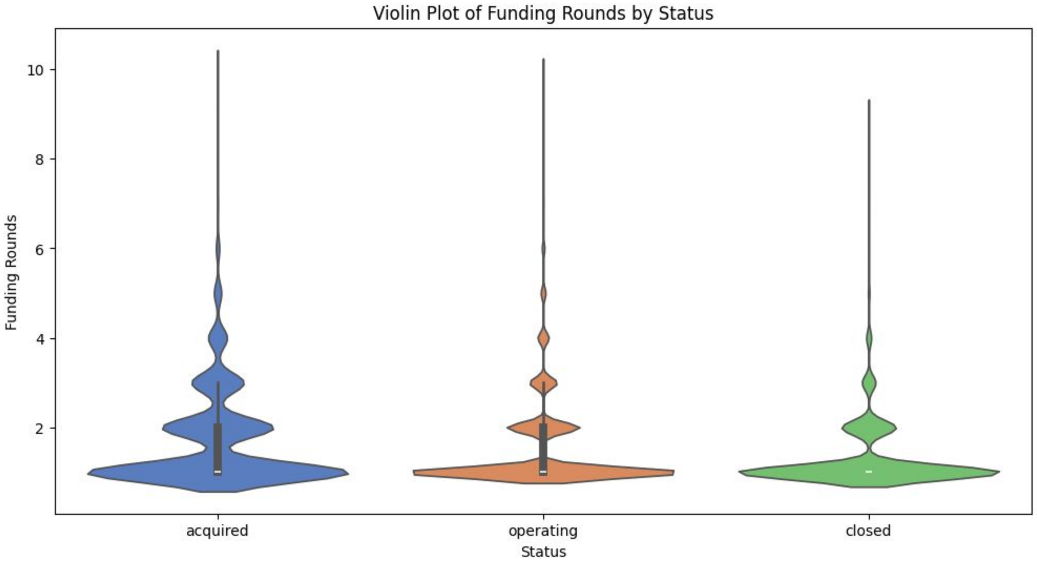
Exploratory Data Analysis - Key Insights

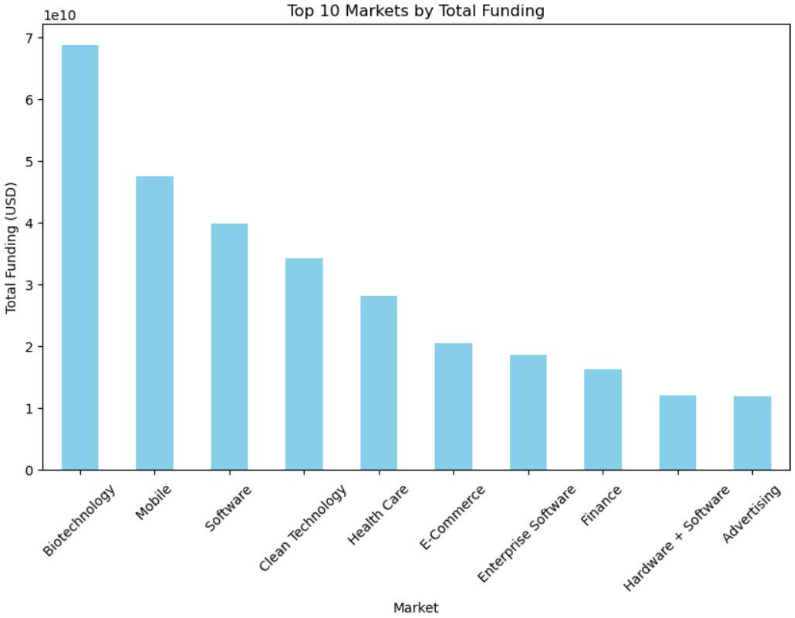


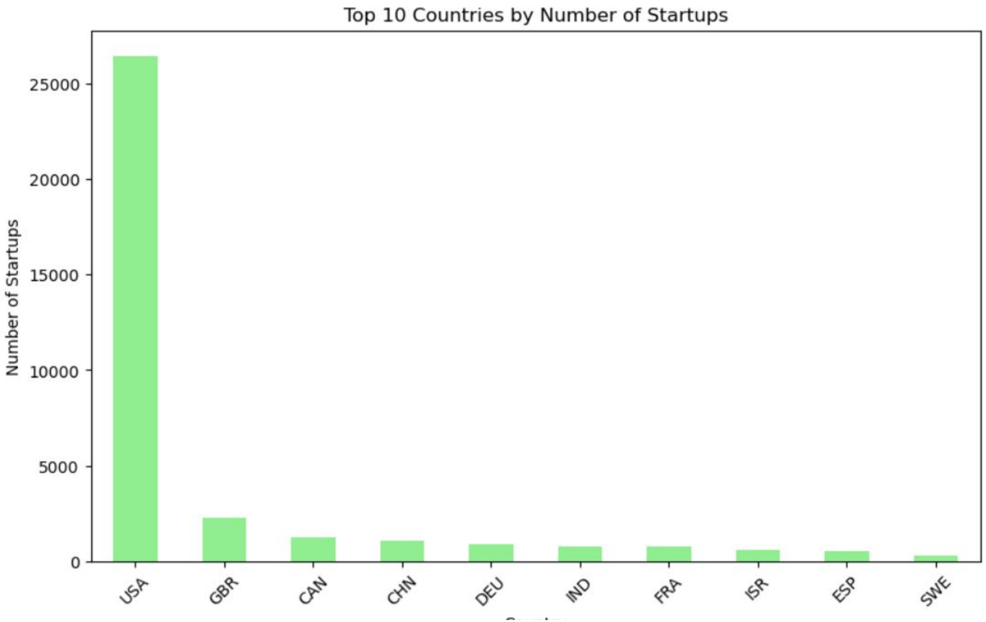
Graph showing the total number of startups year-wise from 2000 to 2014.

The data revealed a noticeable **decline in the number of startups in 2008,** which aligns with the global economic crisis of that year.

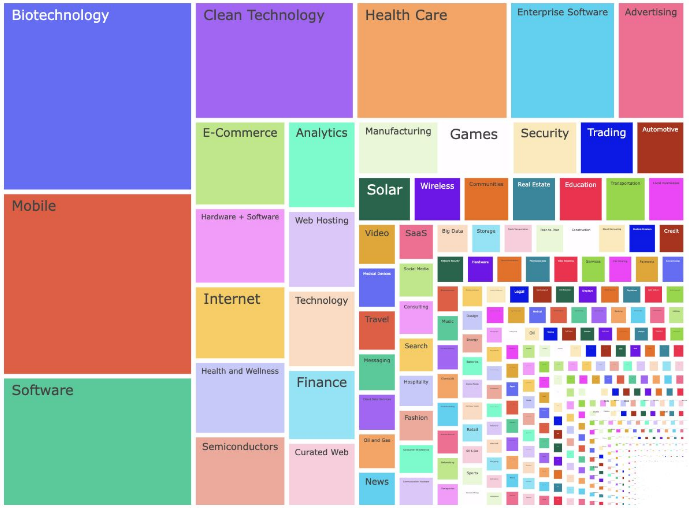
- Following the **recession**, the graph shows a significant rebound, reflecting a period of recovery and renewed growth in the startup ecosystem.

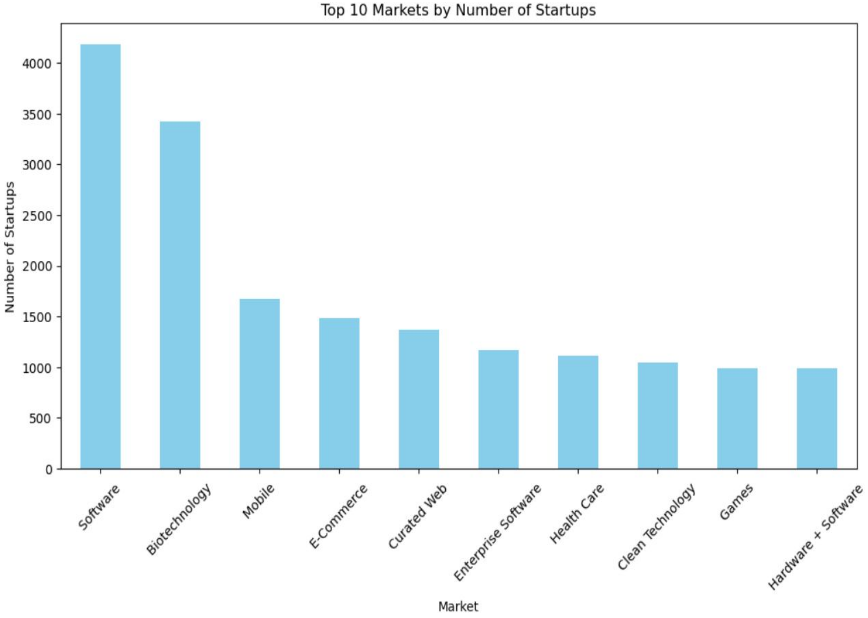

Violin Plot of Funding Rounds by Status
Funding Rounds
10
8
6
4
2
acquired
operating
closed
Status

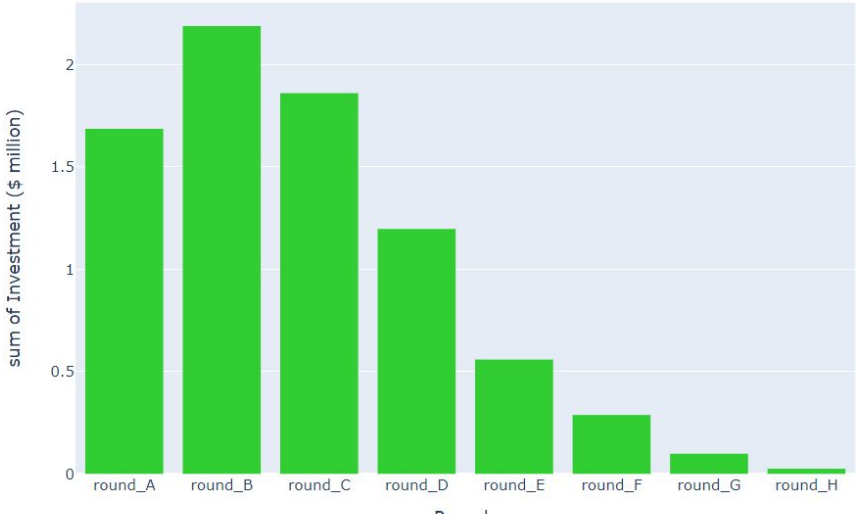

Top 10 Markets by Total Funding
1e10
Total Funding (USD)
7
6
5
4
3
2
1
0
Biotechnology
Mobile
Software
Clean Technology
Health Care
E-Commerce
Enterprise Software
Finance
Hardware + Software
Advertising
Market



Investments made by the USA




Top 10 Markets by Number of Startups
Number of Startups
0
500
1000
1500
2000
2500
3000
3500
4000
Software
Biotechnology
Mobile
E-Commerce
Curated Web
Enterprise Software
Health Care
Clean Technology
Games
Hardware + Software
Market


sum of Investment ($ million)
0
0.5
1
1.5
2
round_A
round_B
round_C
round_D
round_E
round_F
round_G
round_H

Data Preprocessing & Machine Learning Approach

Handled Missing Values:

1. Standardized Date Formats: Converted founded_at, first_funding_at, and last_funding_at to a consistent date format.
2. **Dropped Unnecessary Columns:** Removed irrelevant columns like permalink, homepage_url, category_list, etc.
3. Dropped NaN Rows in 'Status': Eliminated rows where the status column had missing values.

Missing Value Imputation:

4. Imputed ***founded_year and founded_at by grouping data by the market column and filling missing values with median values.***

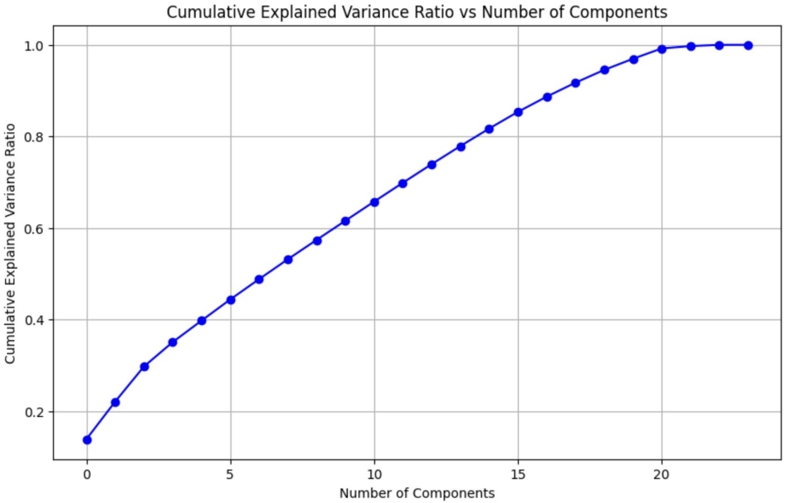
```
DataFrame shape after dropping rows with missing values: (39716, 32)
```

Converted Status to Binary Categories:

Mapped operating and acquired as **Active**, and closed as Inactive.

5. **Did oversampling** such that one class does not dominates.
6. Worked with various ml model like ***Logistic regression, decision tree, random forest.***

Principal Component Analysis



The PCA results show that the first five components collectively **explain 40.0% of the variance in the dataset, highlighting their significance in capturing the underlying patterns related to startup status.** The first component alone accounts for **13.87%** of the total variance, indicating its primary role in the dataset's variability.

```
PCA results:
Number of components: 24
Explained variance ratio of first 5 components: [0.13873226 0.08203882 0.07667265 0.05340594 0.04678652]
```

Deep Dive

Logistic Regression with & without Oversampling

Model	Oversampling	Accuracy	AUC-ROC	Precision (0)	Recall (0)	F1-Score (0)
Logistic Regression (No Reg)	No	0.94	0.52	0.10	0.01	0.02
Logistic Regression (L1 Reg)	No	0.95	0.65	0.20	0.01	0.02
Logistic Regression (L2 Reg)	No	0.95	0.69	0.12	0.00	0.01
Logistic Regression (No Reg)	Yes	0.97	0.99	0.90	0.99	0.95
Logistic Regression (L1 Reg)	Yes	0.99	0.99	0.98	0.99	0.99
Logistic Regression (L2 Reg)	Yes	0.96	0.99	0.89	0.99	0.94

Insights from Logistic Regression

Oversampling is Essential

- Without oversampling, the model heavily favored the majority class, resulting in poor performance on the minority class.

Impact of Regularization

- Both L1 and L2 regularization led to improvements, but L1 regularization consistently showed slightly better precision and recall for the minority class, especially when combined with oversampling.

Near-Perfect Results with Oversampling

- After oversampling, Logistic Regression with or without regularization achieved near-perfect precision, recall, and AUC-ROC, demonstrating the power of this approach in solving data imbalance problems.

Deep Dive - Random Forest

Detailed explanation of Random Forest

- Initialize Random Forest
- Standardize the data
- Applied PCA
- Set up a grid of hyperparameters to tune, including n_estimators, max_depth, min_samples_split
- Use GridSearchCV to find the best combination of hyperparameters via cross-validation. Train the best model using the optimal parameters found.
- Evaluate performance using metrics such as accuracy, precision, recall, validation score

Its specific application in our project

- **Prediction of Startup Status:** Random Forest is used to predict whether a startup will remain "Active" or "Inactive."
- **Handles Complex Relationships:** The model captures non-linear relationships between features like funding, market growth, venture capital
- **Accurate and Stable Predictions:** By averaging multiple decision trees, Random Forest improves prediction accuracy and reduces overfitting.
- **Feature Importance:** It helps identify key factors influencing startup success, aiding stakeholders in data-driven decision-making.

Results and Evaluation

- Key performance metrics

Best hyper-parameter found:

```
Decision Tree Results:
              precision    recall  f1-score   support

           0       0.12      0.04      0.06       436
           1       0.95      0.98      0.96      7508

    accuracy                           0.93      7944
   macro avg       0.53      0.51      0.51      7944
weighted avg       0.90      0.93      0.91      7944

Accuracy: 0.9312688821752266
AUC-ROC: 0.5121179610250595
```

```
Best Parameters:  {'n_estimators': 100, 'min_samples_split': 2, 'max_depth': 10}
```

```
Cross-Validation Scores: [0.93764151 0.93773585 0.93632075]
Mean Cross-Validation Score: 0.9372
Test Accuracy: 0.9357
F1 Score: 0.9149
Precision: 0.8960
Recall: 0.9357
```

Challenges Overcome During Model Development:

Imbalanced Data: The dataset had more "Active" startups than "Inactive" ones, leading to biased predictions. We used oversampling and downsampling to have better ratio

Feature Selection: Some features were highly correlated or irrelevant, which affected the model's performance. We used feature importance and correlation analysis to remove redundant features.

Hyperparameter Tuning: Finding the right hyperparameters was challenging and also the computational cost was high and time taken process , so we applied GridSearchCV to fine-tune parameters like n_estimators, max_depth

Real World Applications and Impact

Potential Use Cases

- **Investment Decisions**: Identify high-potential sectors and locations for funding.
- **Market Analysis**: Analyze startup trends to forecast market shifts.
- **Policy Development**: Inform government initiatives to support entrepreneurship.
- **Networking**: Facilitate partnerships between startups in similar markets.

Implementation

- **Data Integration**: Merge startup data with analytics platforms for real-time insights.
- **Predictive Analytics**: Use machine learning to forecast startup success.

Estimated Impact

- **Cost Savings**: Reduce investment risks, saving capital.
- **Efficiency Gains**: Accelerate decision-making processes.
- **Informed Policies**: Enhance resource allocation for entrepreneurial support.

Future work and conclusion

Areas for Improvement or Expansion

- **Data Enrichment**: Add more datasets for deeper insights.
- **User Customization**: Allow users to tailor dashboards to their needs.

Next Steps

- **Beta Testing**: Launch a pilot version for user feedback.
- **Partnership Development**: Collaborate with industry players for validation.
- **Ongoing Research**: Analyze long-term trends in startup success.

Empowering Entrepreneurs: Data-driven insights can enhance innovation and drive economic growth.

Thank You!

Questions?