



PROSPECT TO PROSPERITY!

Unraveling Startup Success with Machine Learning Model Predictions

Here's what we've found



TEAM 9B

HUSAIN MANDLIWALA

NIKHIL KOKA

PEIQI HUANG


TANAYA RANADE

WEI-SHIANG WANG



INTRODUCTION

WHY DOES THIS NEED TO BE STUDIED?

- According to research by the U.S. Bureau of Labor Statistics, the failure rate for startups is
 -  20% within first 2 years
 -  45% within first 5 years
 -  65% within first 10 years
- Highly competitive landscape for new entrepreneurs
 - Market Dynamics
 - Capital Constraints
 - Executive Mismanagement & more
- Yet, success is not impossible...

1361

unicorns in 2023
worldwide

\$221bn

global vc funding in
2023 for startups

WHAT NEEDS TO BE STUDIED?

EVENTS LEADING TO 'SUCCESS'

- Merger & Acquisition
- Startup > Unicorn
- IPO Launch

PROBLEM STATEMENT DEVELOPMENT

- Which machine learning model can best predict whether a startup will be successful or not?
- How can the machine learning models predicting startup success be optimized for most accurate performance?
- Which factors are important to consider while predicting the success of a startup?

DATASET DESCRIPTION

Source: Kaggle

922

unique records

49

attributes

Data Cleaning

- Dealing with Missing & Null Values
- Reformatting Data (e.g. into DateType object)
- Dummy Encoding
- Removing Irrelevant Variables

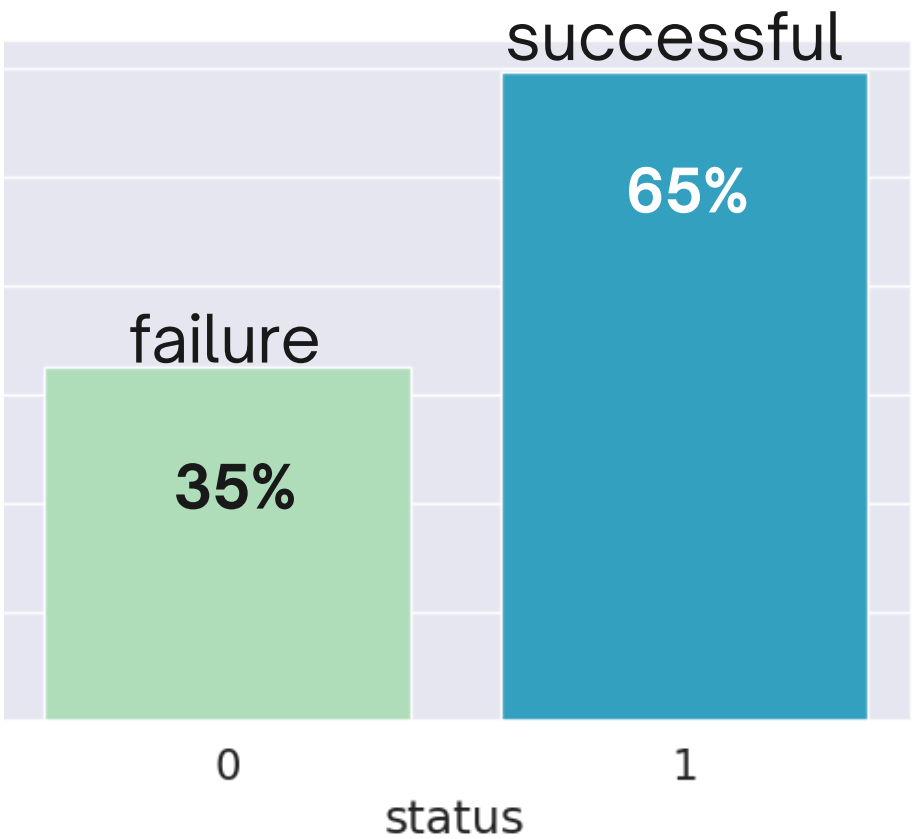
922

unique records

302

attributes

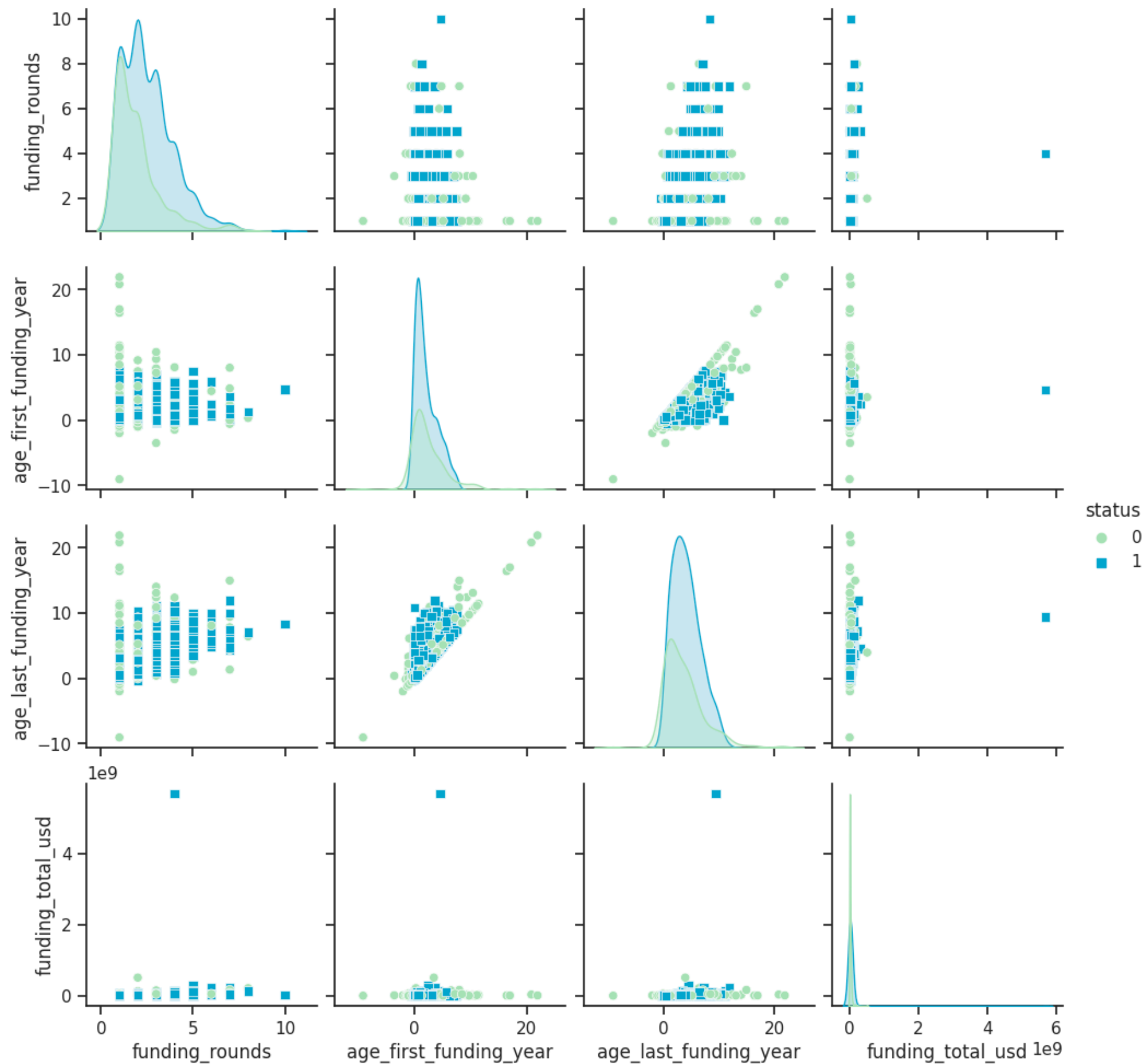
Target Variable | Class Distribution



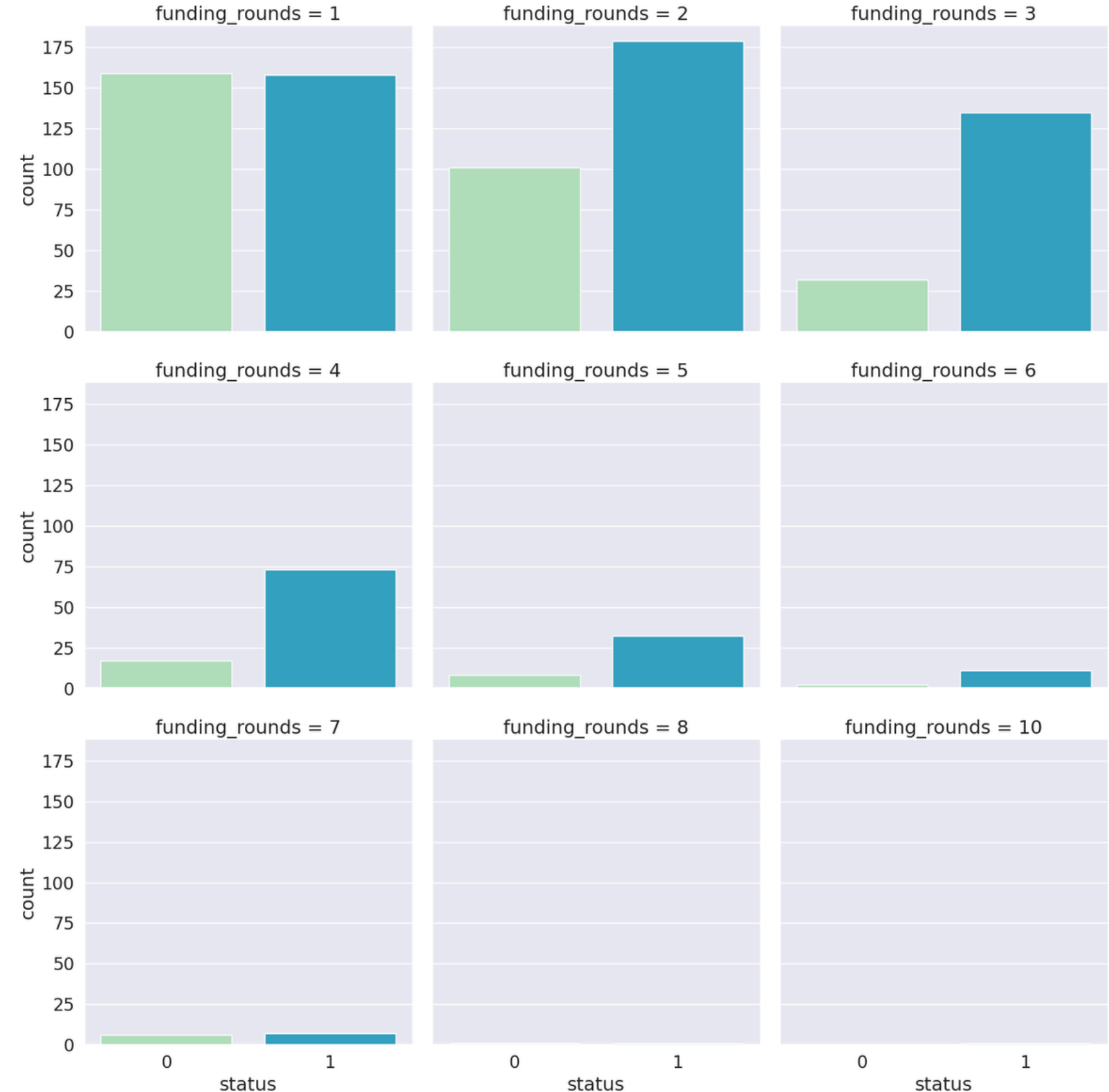
Key Independent Variables

total funding	industry category	founding city
funding rounds	is in top 500	founding state
avg. employees	age in first & last funding year	age in first & last milestone

EXPLORATORY DATA ANALYSIS



Overview of pairwise relationships between numeric variables



Is the number of funding rounds relevant for evaluating startup success?

EXPLORATORY DATA ANALYSIS



Age of startups during their first funding vs last funding
Did this impact success?

Failed Startups

First funding @ 0-2 years
Last funding @ 0-5 years

Successful Startups

First funding @ 0-4 years
Last funding @ 1-10 years

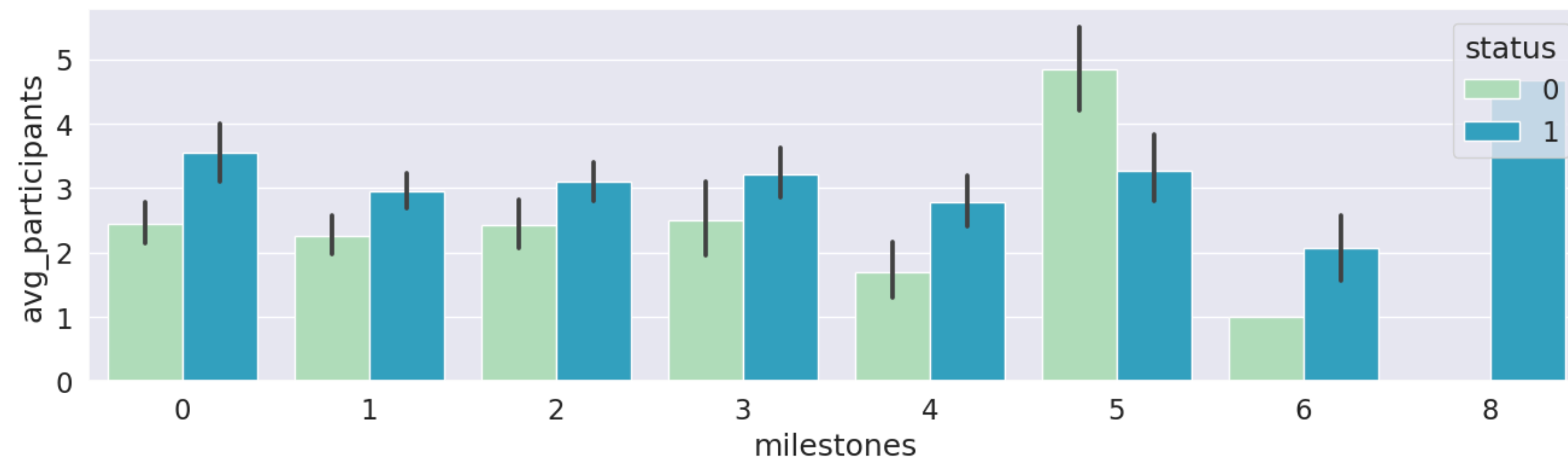
*Does a bigger company
always imply more success?*

Successful Startups

Stable organization size with growth

Failed Startups

Erratic org size with growth



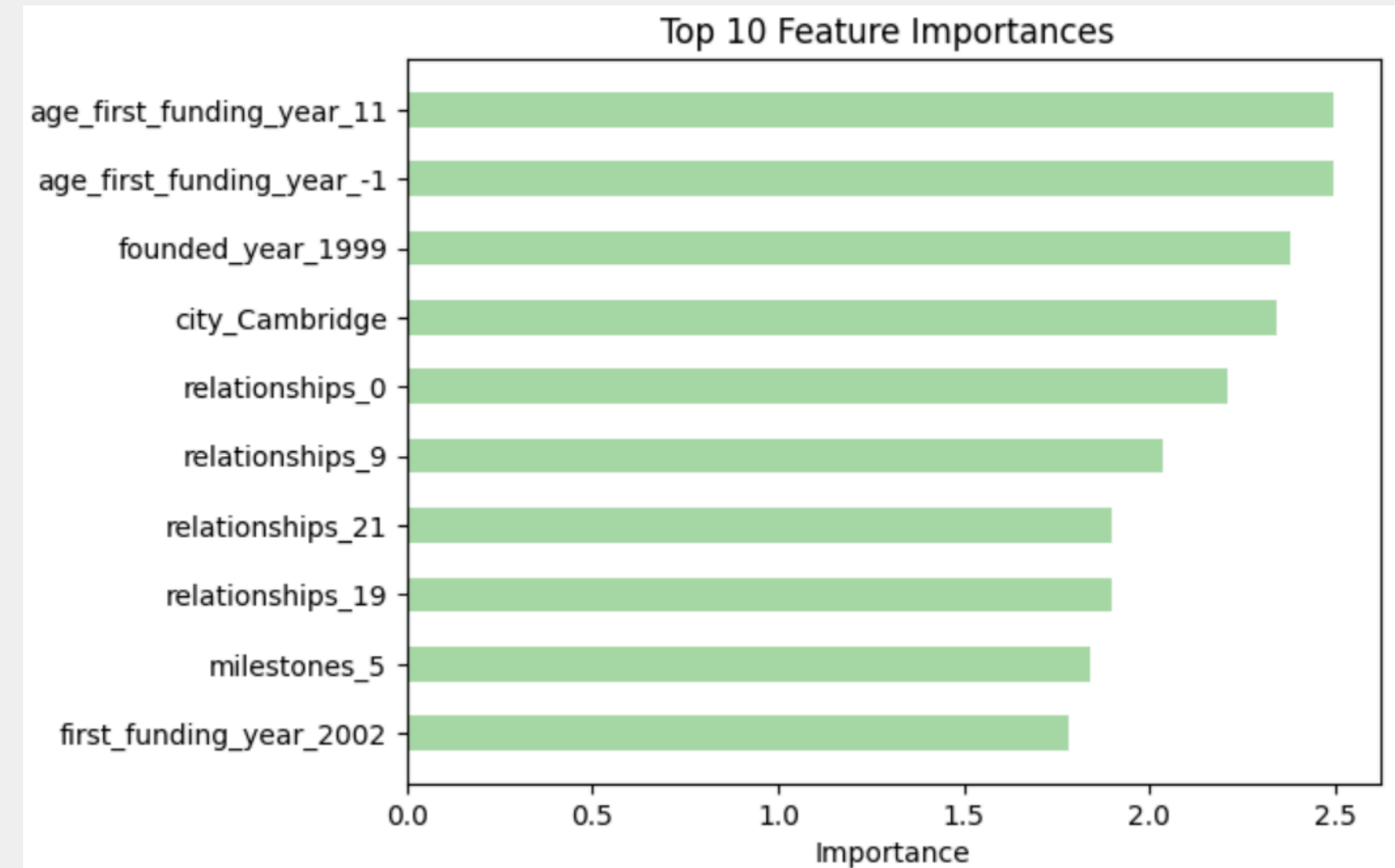
NAIVE BAYES MODEL

INTRA-MODEL PERFORMANCE EVALUATION

Model	Accuracy	Accuracy (Success)	Accuracy (Failure)	F1 Score
Unprocessed	79%	86%	67%	80%
Cross Validation	78%	84%	65%	84%
SMOTE	69%	78%	67%	80%

- Neither cross-validation nor SMOTE has significantly improved the accuracy or reliability of this model
- The one derived post stratified KFold cross-validation provides the best performance in terms of accuracy as well as F1 measure

TOP 10 FEATURE IMPORTANCES

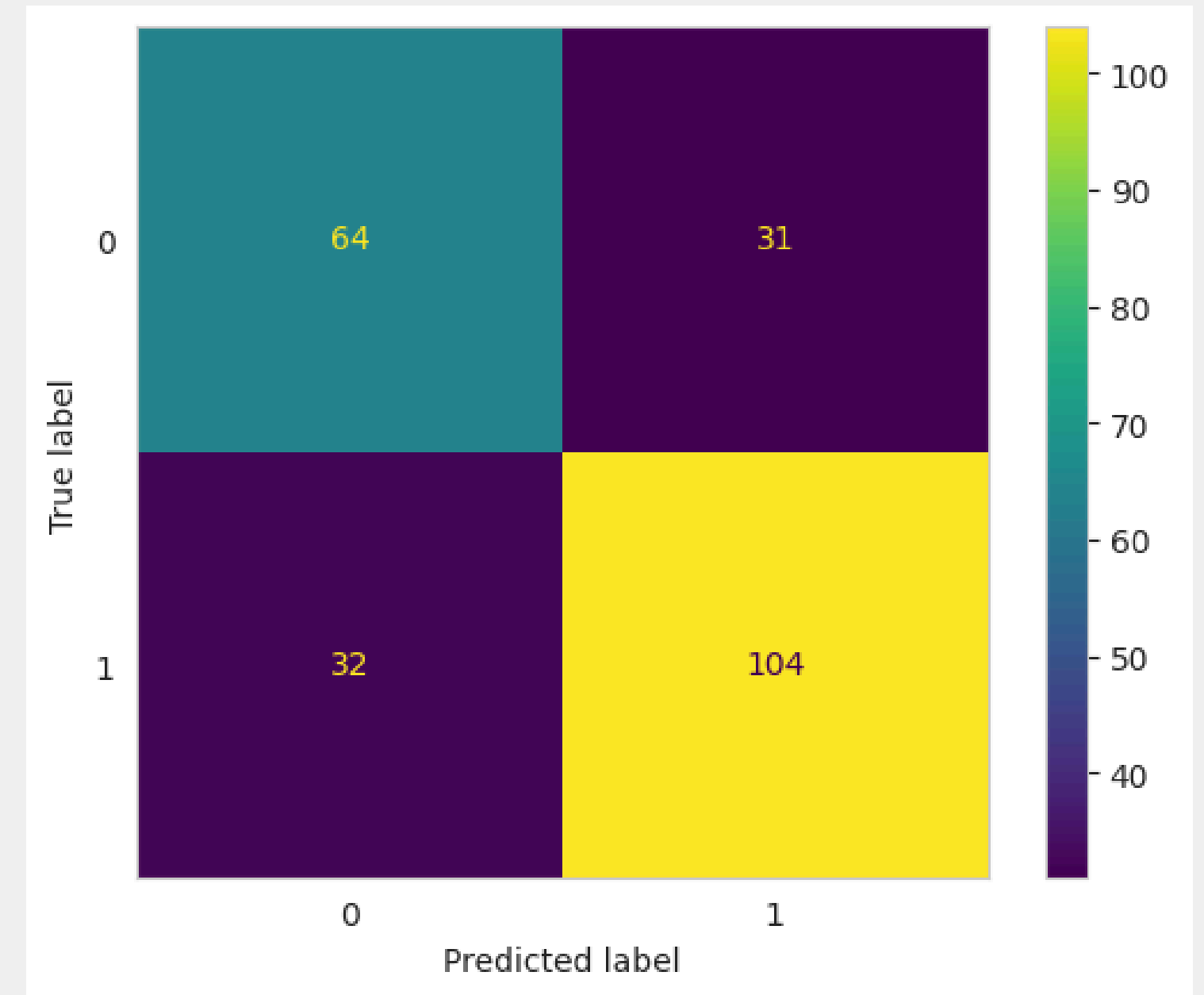


- The time of founding of the startup and the age when it receives its first funding are crucial factors in determining startup success

DECISION TREE MODEL

Model	Accuracy	Accuracy (Success)	Accuracy (Failure)	F1 Score
Unprocessed	82%	95%	60%	83%
Post Pruning	73%	75%	69%	73%
Grid Search	72%	76%	67%	72%

- Unprocessed model - cleaned data
- Post Pruning
- Grid Search
 - Gave us the best accuracy overall

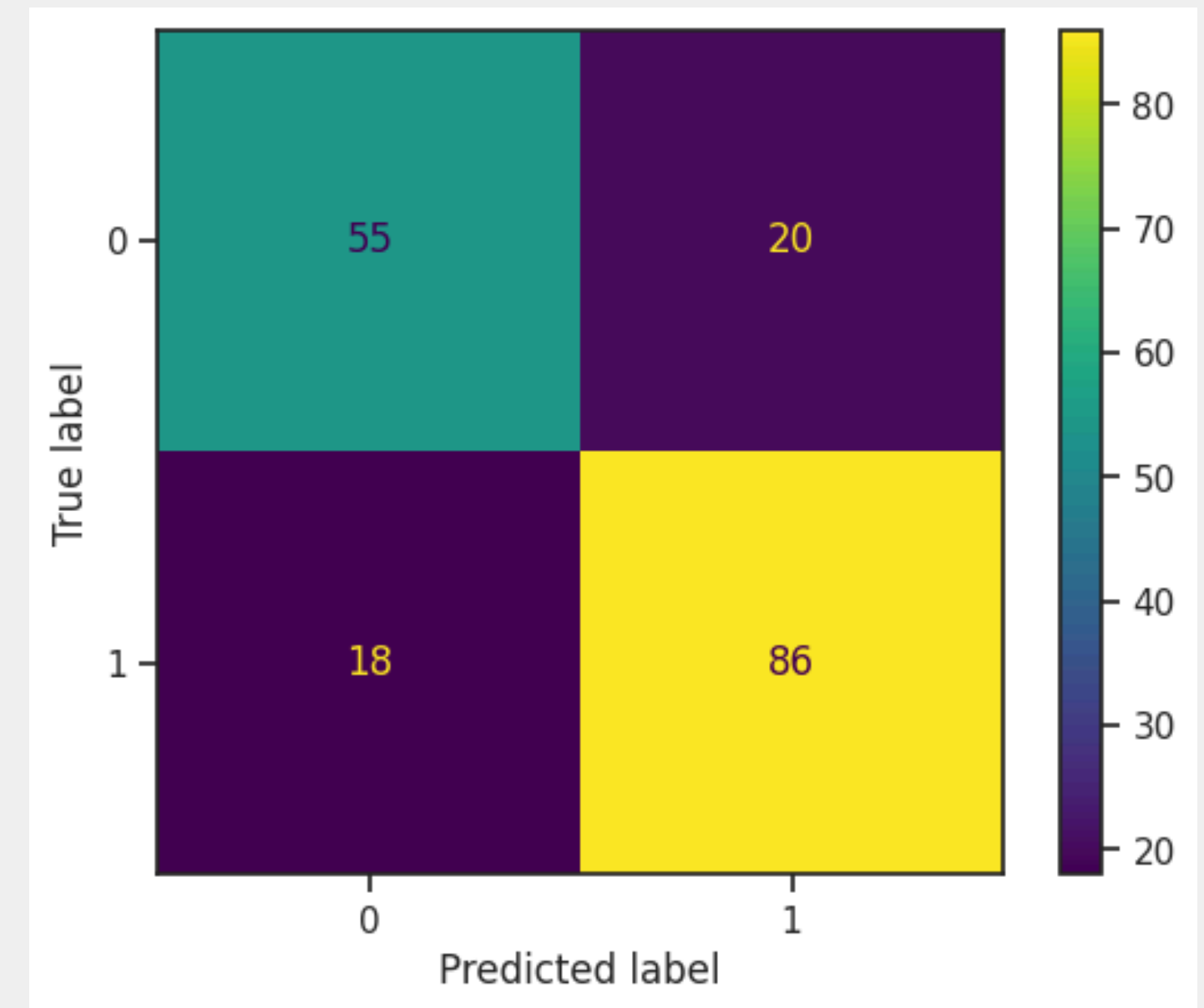


Too many categorical variables to slit
Because of which chance of over fitting the data is
very High

LOGISTIC REGRESSION MODEL

Model	Accuracy	Accuracy (Success)	Accuracy (Failure)	F1 Score
Unprocessed	64%	100%	0%	78%
Grid Search Cross Validation	75.8%	85.31%	59%	76%
SMOTE	78.7%	82.7%	73.3%	78.8%

- Unprocessed model - cleaned data
- Grid Search CV yielded parameters:
 - Penalty = l1 / lasso regression
 - C = 1
- SMOTE model - 70-30 sampling strategy
- Best Intra-model - SMOTE



SMOTE was best model because:

- Most optimum performance on basis of accuracy, stratified accuracy & reliability (F1)
- Well-balanced confusion matrix ; reduced overfitting from previous models

INTER-MODEL EVALUATION

NAIVE BAYES

- While Accuracy & F1 were good, model was overfitting data, as seen from confusion matrix.

DECISION TREE

- Similarly, Accuracy & F1 were good, but decision tree is not a good choice for the following data set - high dimensionality.

LOGISTIC REGRESSION

- Provided the most optimum scores on Accuracy & F1
- Reported best stratified acc.
- Resolved overfitting issues

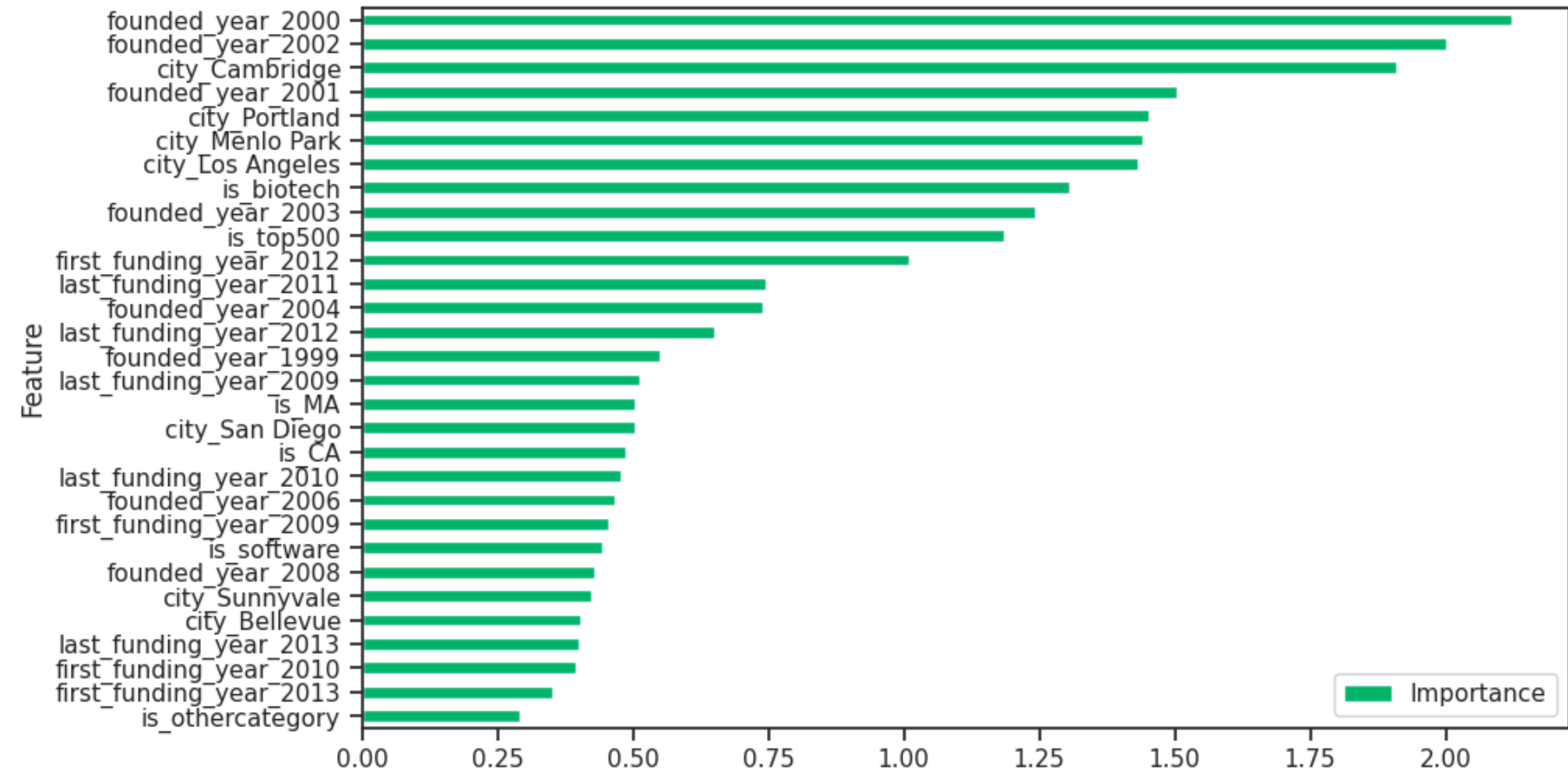
Model	Accuracy	Accuracy (Success)	Accuracy (Failure)	F1 Score
Naive Bayes Classifier	78%	84%	65%	84%
Decision Tree Model	72%	76%	67%	72%
Logistic Regression Model	78.7%	82.7%	73.3%	78.8%

INSIGHTS & IMPLICATIONS

MOST IMPORTANT FEATURES

as per logistic regression model

- Founded between 2000 & 2003
 - Executives can study market conditions, funding trends, etc. & try to adapt for success in current times
- City of operations
 - Executives can study city size, demographics, market preferences & business opportunities to capitalize on similar factors or relocate
- Biotechnology industry
- Top 500 companies



DECISION TREE

>

NAIVE BAYES

>

LOGISTIC REGRESSION

THANK YOU!

and we hope you found our project interesting!