Doomed to Succeed? Operation Agnostic Data and Startup Success

FNCE237 Data Science for Finance
Group 1 | Homer Wang, Jamie Lee, Robert Dowling

Background & Data

Are there permanent features of a startup, that have nothing to do with their operation, that predict success of the firm?

Problem

Our group set out to create a prediction tool for Venture Capitalists to assess startups for potential financing.

Instead of evaluating dynamic features, such as revenue, customers, and operations, we assessed static, permanent pieces of information that VCs may usually find irrelevant.

They include, but are not limited to, the number of syllables in the name of the company, the company's geographic location, the sentiment scores of the company description, etc.

Data

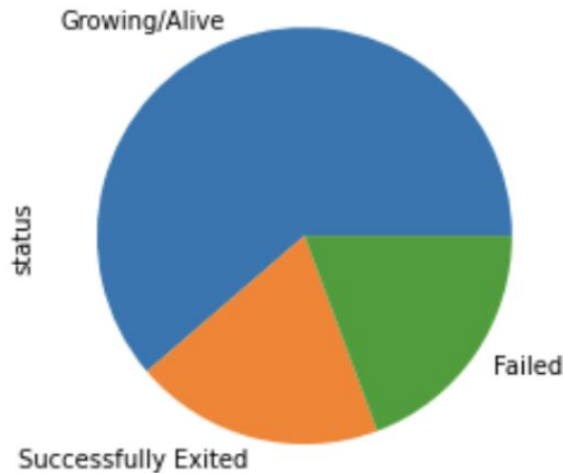The data we used came from Pitchbook, thanks to WR&DS.

Pitchbook provides data on startups that are seaking financing.

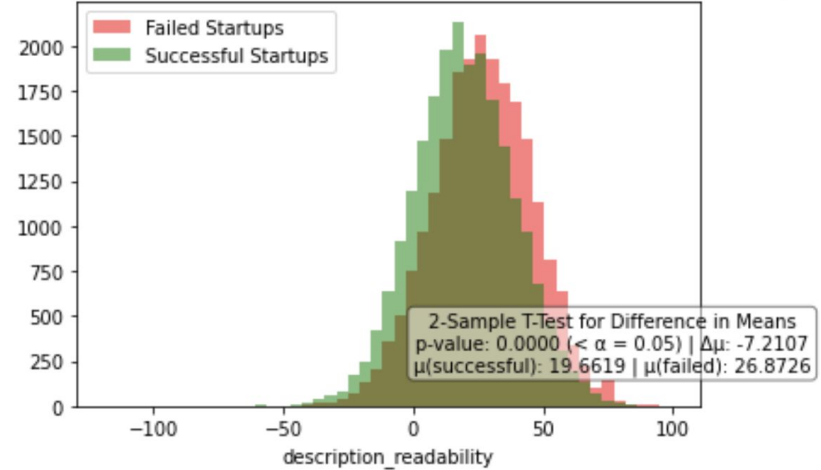We chose 107 different parameters by which to assess the businesses.

Select Initial Exploratory Data Visualizations

Initial inspection of the data provided insight into how we defined startup success and the parameters we would use.

Defining Startup Success

A success is a firm that IPOs or is acquired.
A failure is a company that is dissolved.
All growing companies were rejected for this model.

Example Parameter

For each firm,
107 parameters were assessed.
Plots demonstrating whether more startups succeeded or failed and given parameter values demonstrates unexpected trends.
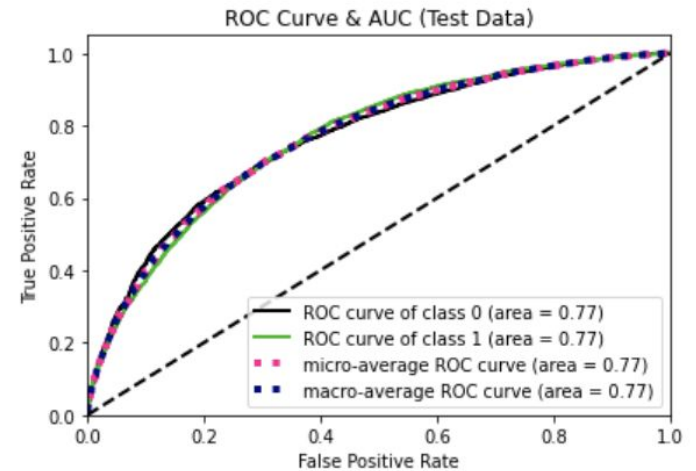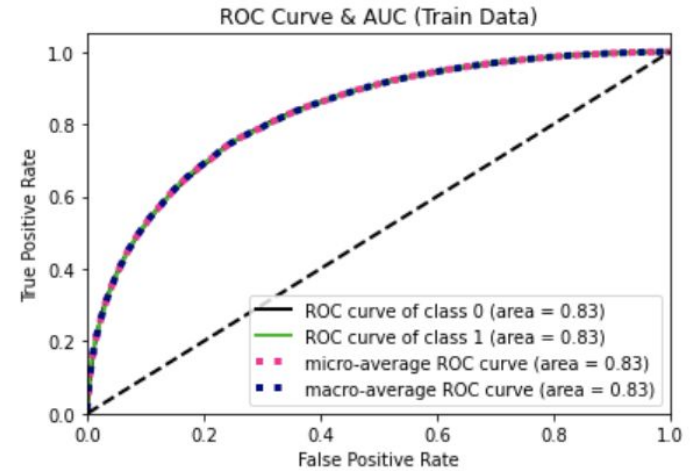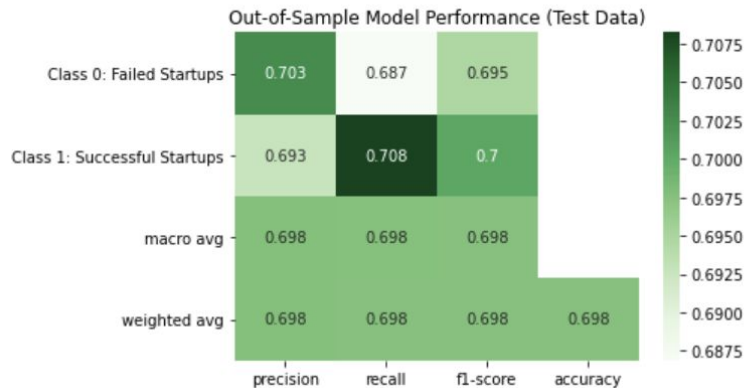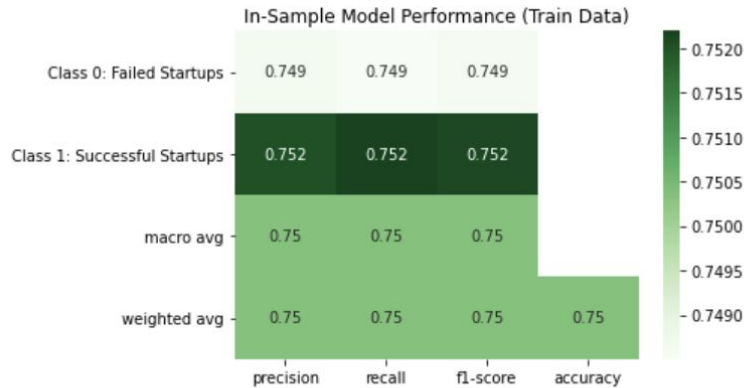
Results

Our XG Booster Classifier, with hyperparameter tuning, performed relatively well compared to the training data.

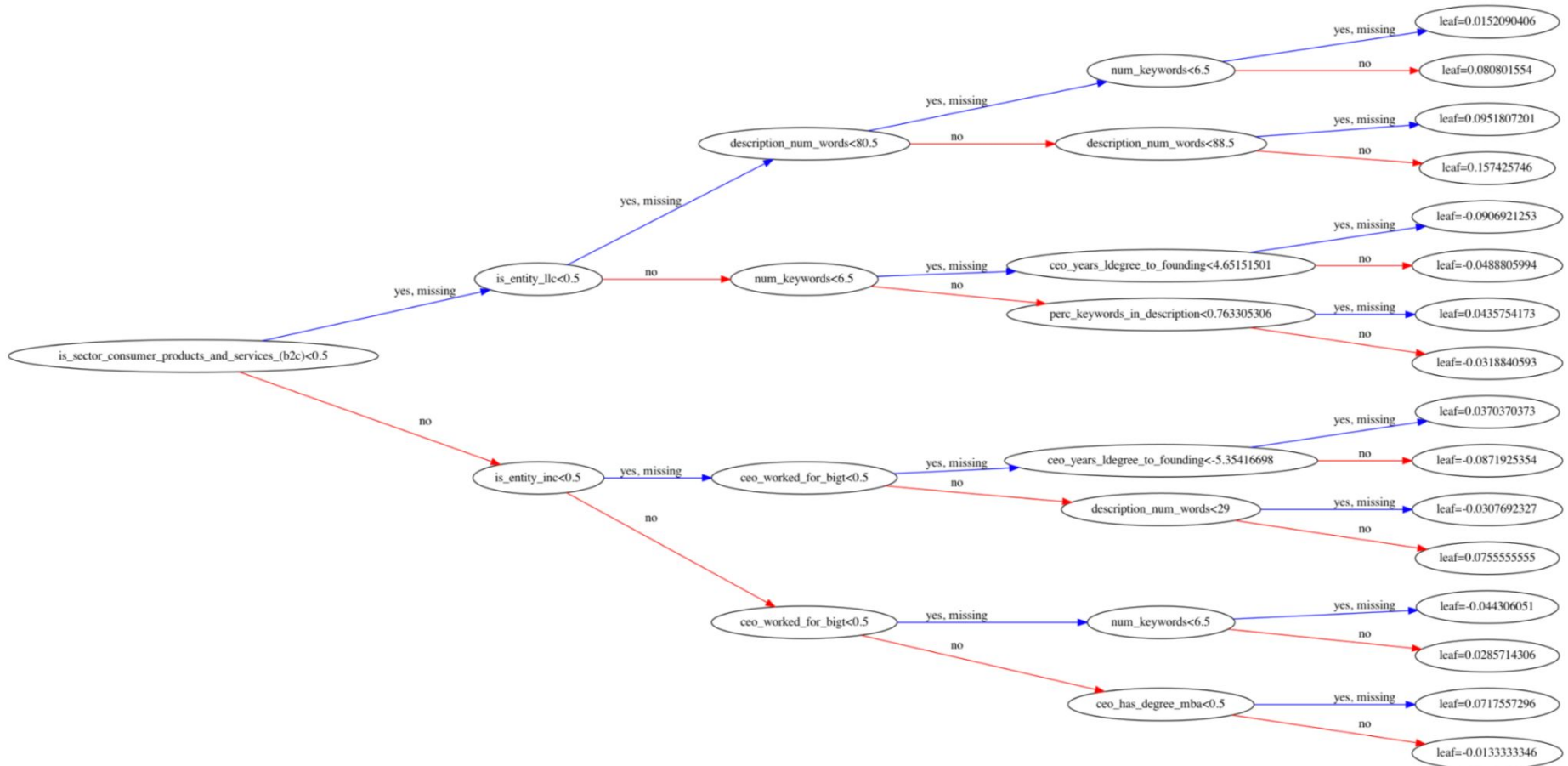====================================== Model: XGBClassifier ======

Cross Validation Results
Different predictions: 3,210 (9.09%)
CV avg model accuracy score: 69.25%
CV min model accuracy scores: 68.70%
CV max model accuracy scores: 69.81%

In-Sample Model Performance (Train Data)

	precision	recall	f1-score	accuracy
Class 0: Failed Startups	0.749	0.749	0.749	
Class 1: Successful Startups	0.752	0.752	0.752	
macro avg	0.75	0.75	0.75	
weighted avg	0.75	0.75	0.75	0.75

Out-of-Sample Model Performance (Test Data)

	precision	recall	f1-score	accuracy
Class 0: Failed Startups	0.703	0.687	0.695	
Class 1: Successful Startups	0.693	0.708	0.7	
macro avg	0.698	0.698	0.698	
weighted avg	0.698	0.698	0.698	0.698

ROC Curve & AUC (Train Data)

ROC curve of class 0 (area = 0.83)
ROC curve of class 1 (area = 0.83)
micro-average ROC curve (area = 0.83)
macro-average ROC curve (area = 0.83)

ROC Curve & AUC (Test Data)

ROC curve of class 0 (area = 0.77)
ROC curve of class 1 (area = 0.77)
micro-average ROC curve (area = 0.77)
macro-average ROC curve (area = 0.77)

Tree Branch Visualization

The model "thinks" about the relationship between parameters by weighing 382 "trees." Here is a visualization:

Thank you!