Project Proxy - Predicting Startup Success

Raymond Zhao
Yigit Ihlamur

July 24, 2022

Abstract

Previous iterations of the project concluded successful investors signal successful startups. This report aims to explore this proxy and identify signals that distinguish between companies invested by failed and successful investors, and between companies invested by failed and brand investors. Through feature engineering, statistical analysis, and ML models like SVC and GBC, we find that the most important attributes are Founder's University Ranking and Employer Reputation, Founder's previous experience in start ups, and geographical location (specifically San Francisco, we subsequently explore how this is affected by Covid). The code can be found at https://github.com/raymondyqz/Proxy

1 Introduction

Predicting future outcomes in early-stage investing is more difficult than other investment fields. There are less numbers and more qualitative features. With the proliferation of data about private companies and advances in machine learning, we can build models to predict the future performance of a startup much better than relying on our intuitions. Prior work at Vela proved that investors of a startup are the most important predictive feature of future success. Our thesis is that this feature is a critical proxy signal that we should deep dive and understand the correlation of other attributes deeper. In this project, our goal is to build a model to predict if successful investors are going to invest in a startup.

2 Data

We are given 3 datasets which detail attributes of companies invested by failed, successful and brand investors respectively. Based on previous quantitative analysis, successful investors have a history of investments that beats a if you were to invest in startups randomly-can interpret as 'market average', while failed investors have success rate that is lower than the market average. Brand investors are the ones that have a well-known brand in the market (most of them are also successful investors). The datasets are an extraction of investments from a subjective list of investors. Each dataset consists of 3 data sheets: List, Academic and Work. The 'List' sheet has companies with their main attributes. The 'Academic' sheet contains a mostly academic background of the founders. The 'Work' sheet lists the companies that the founders worked before and what their titles were. The organization name is the unique field that can be used to connect the 3 data sheets. For each dataset, we merge the data sheets into a single data frame, only including the companies which have data in all 3 data sheets. Then, we merge the data frames for the 3 datasets. To distinguish between the data, we add the response variable 'success_flag', which is set to be 0 for companies invested by a failed investor (4989 instances), 1 for companies invested by a successful investor (2030 instances), and 2 for companies invested by a brand investor (1430 instances). There are 8449 organizations in total. The explanatory variables that we will consider are shown below:

- 'city': The city that the organization is based on.
- 'city_of_founders': The cities founders are from.
- 'universities_of_founders': The universities attended by the founders.

- ‘degrees_of_founders’: The university degrees obtained by the founders.
- ‘subject_degrees_of_founders’: The subjects undertaken by the founders in university.
- ‘gender_of_founders’: The genders of the founders.
- ‘prev_companies_of_founders’: The companies that the founders worked before.
- ’founded_on’: Date company was founded

We also use the following datasets to help us generate more features:

- ‘2022_QS_World_University_Rankings_Results_public_version_modified.xlsx’
- ‘unicorn.csv’: List of companies that were founded after 2008, raised a minimum of $300M, went for an IPO (initial public offering) and are based in the US.
- ’Startups_per_stage_based_on_total_funding.csv’: List of all startups since 2008 categorised into stages based on total funding.

3 Feature Generation and Analysis

3.1 Feature Generation

As most features in the data-set are categorical, we generate numerical features by:

1. Flagging the inclusion of desired attributes (i.e. a certain degree, university).
2. Calculate the percentage of that desired attribute occurring(i.e. percentage of female founders).
3. For Ordinal Data, we create standardised numerical scores (i.e. for University degrees, we give Undergraduate a 1, Masters a 2, and Doctoral a 4)
4. University Scores (20-100) and University Employer Reputation Scores (1-100) are obtained from the QS World University Rankings 2022 dataset and normalised.

The following are the features initially generated:

- ’success_flag’
- ’maximum_founders_university_score’
- ’minimum_founders_university_score’
- ’average_founders_university_score’
- ’num_universities’
- ’number_of_founders’
- ’number_of_male_founders’
- ’number_of_female_founders’
- ’percentage_of_male_founders’
- ’degrees_of_founders_standardised’
- ’degrees_of_founders_standardised_numbers’
- ’best_degree_of_founders’
- ’worse_degree_of_founders’
- ’average_degree_of_founders’

- ’percentage_of_founders_with_a_degree’
- ’in_sf’: Flag is company is based in San Francisco - this was put in intuitively due to Silicon Valley being located in San Francisco.
- ’percentage_founders_in_sf’
- ’big_flag’: Flag if founders have past experience in big tech companies
- ’big_percentage’: Percentage of founders with experience in big tech
- ’unicorn_flag’: Flag if founders have past experience in Unicorn companies
- ’unicorn_percentage’: Percentage of founders that have past experience in Unicorn companies

3.2 Feature Analysis and Feature Importance Ranking

We focus on 4 particular categories: Education, Past experience, Geography, and Gender. Using the 2 datasets for successful vs failed and brand vs failed, we ranked the correlation of different features to ’success_flag’ to test feature importance. Note features like university qualification has no correlation to success, while having female founders gives slight negative correlation (can be due to discrimination and lack of diversity).

Table listing top 5 features (by correlation to success flag) for successful vs failed:

Feature	Correlation to ’success_flag’
’percentage_founders_in_sf’	0.1866
’in_sf’	0.1759
’maximum_founders_university_score’	0.1303
’average_founders_university_score’	0.1260
’unicorn_percentage’	0.0886

Table 1: Top features by importance (correlation to success).

The box plots for some of the top features for brand vs failed:

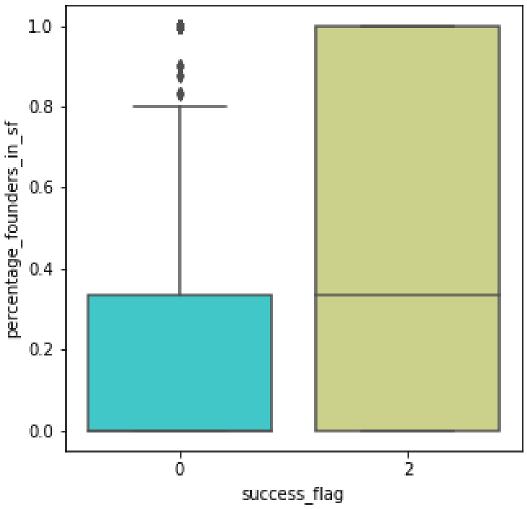


Figure 1: Founders in San Francisco

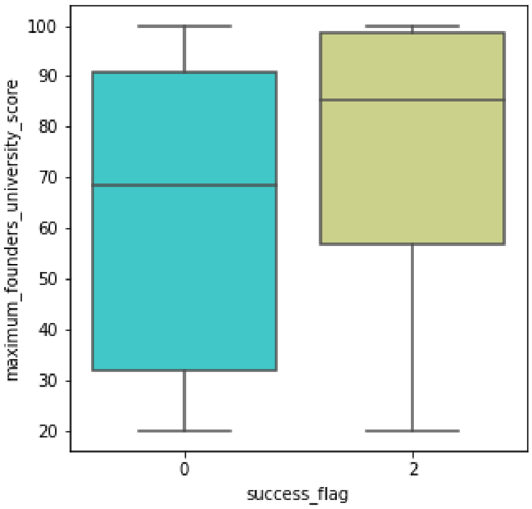


Figure 2: Max Founder’s University Score

3.3 Feature Engineering

From the feature importance rankings, we pick up a few important signals, namely:

- University founders went to.
- Past experience in Start-ups, as opposed to in big tech companies.
- Geography- of both founders and the company

And hence we engineer sub-features and relevant features to break down the signals and find more elegant proxies.

For University we created features like university employer reputation, if the founders undertook a stem or technical degree, particular universities (Stanford, MIT, IVY league, top international universities), and the location of their universities (US or international).

Out of those features, and on successful vs failed dataset (trends replicated in brand vs faield), employer reputation has the highest correlation to success (0.1194), then followed by Stanford flag (0.0830), stem flag (0.0753) and IVY flag (0.0729). Attributes like MIT, international universities, Top International Universities have no strong correlation.

For past experience in start ups, we used the dataset 'Startups_per_stage_based_on_total_funding.csv', which pulled startups from 2008 and using feature engineering, defined the financing stage of each company according to this total funding. The stages being seed, no funding, series a, series b, series c, series c or unicorn.

This is a very dynamic set of attributes, founders with past series C or Unicorn experience showed extremely high correlation to success (0.1490), and this correlation decreases with the founding stages, with the lowest being seed round startups, with a negative correlation (-0.1569). This might be explained as follows: founders that were successful in their past startups (i.e. became Unicorn) tend to continue being successful, while founders with a bad track record (only made it to seed round) tend to not be picked by good investors.

Feature	Correlation to 'success_flag'
'c_unicorn_per'	0.1490
'series_c_per'	0.1058
'series_b_per'	0.1148
'series_a_per'	0.0410
'unfunded_per'	0.0424
'seed_per'	-0.1569

Table 2: Percentage of founders with certain start-up expereinces ranked by correlation to success.

For Geography, we tested whether other major cities in the US also have strong correlation to success, and investigated New York and Miami. We have also split the dataset into pre and post Covid to see if the geographical advantages still persist after Covid and the rise of remote working.

While the other locations don't have a big correlation, we can see the relative importance of San Francisco decreasing after Covid as compared to before.

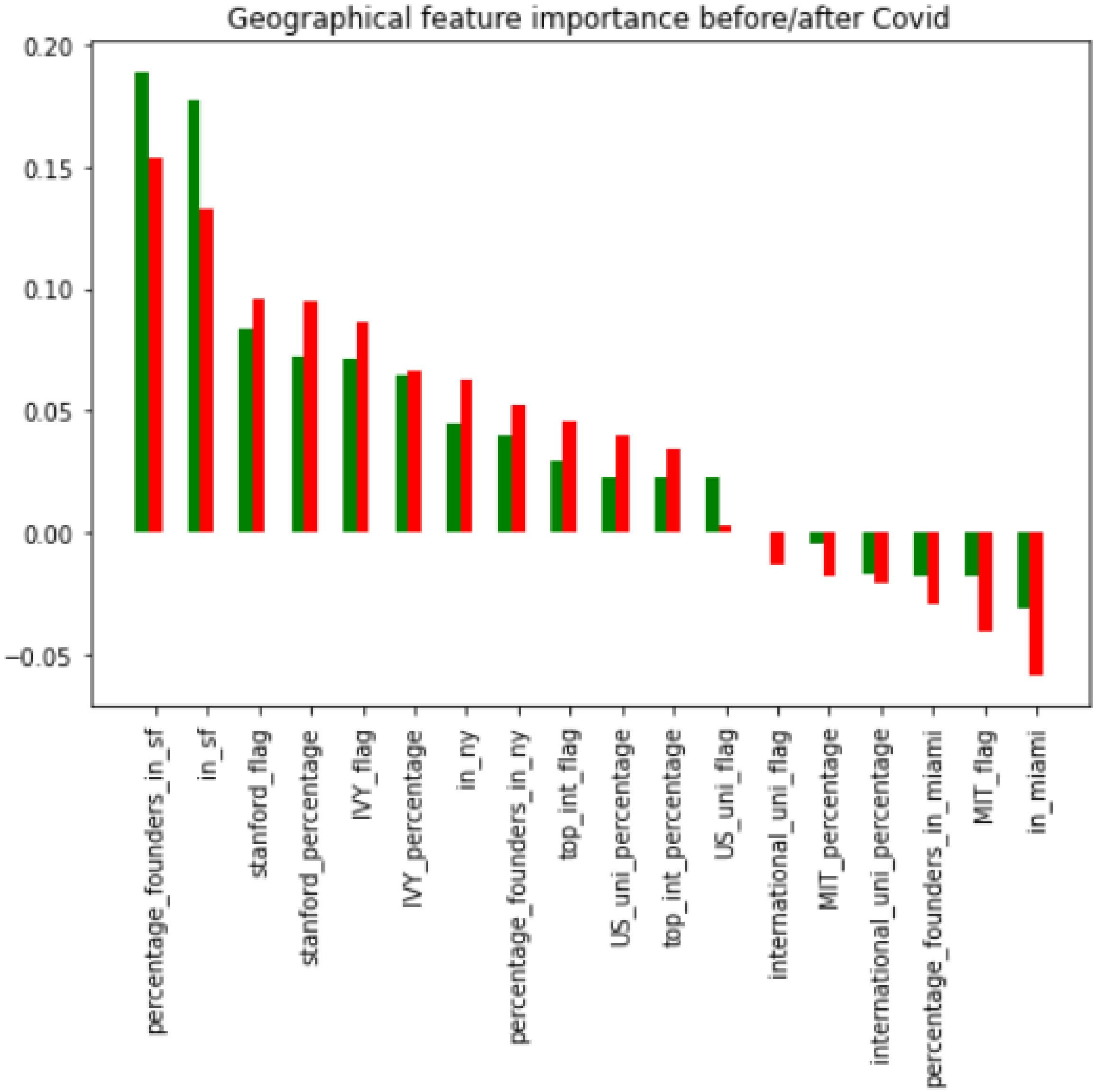


Figure 1: Geography features before (Green) and after(Red) Covid

Below is a figure ranking the Feature Importance of all features:

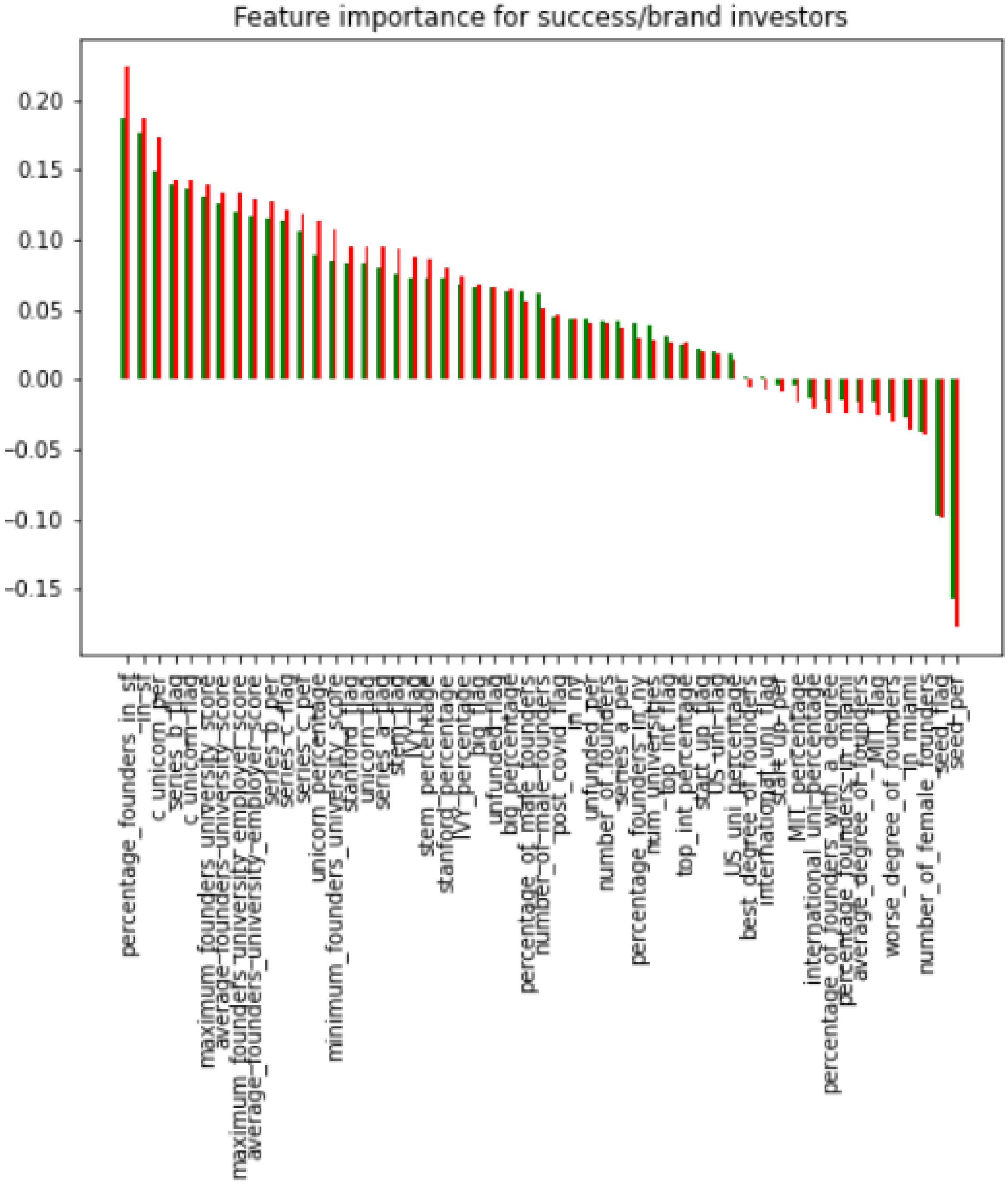


Figure 2: All Feature Importance Ranked

4 Model

4.1 Matrices

As both the combined successful, brand, and failed datasets are highly imbalenced the usual accuracy mectric is not good enough to evalute the machine learing models and features. Therefore we require the precision and recall metrics which are better suited to evaluating the models ability to idetnify True positives

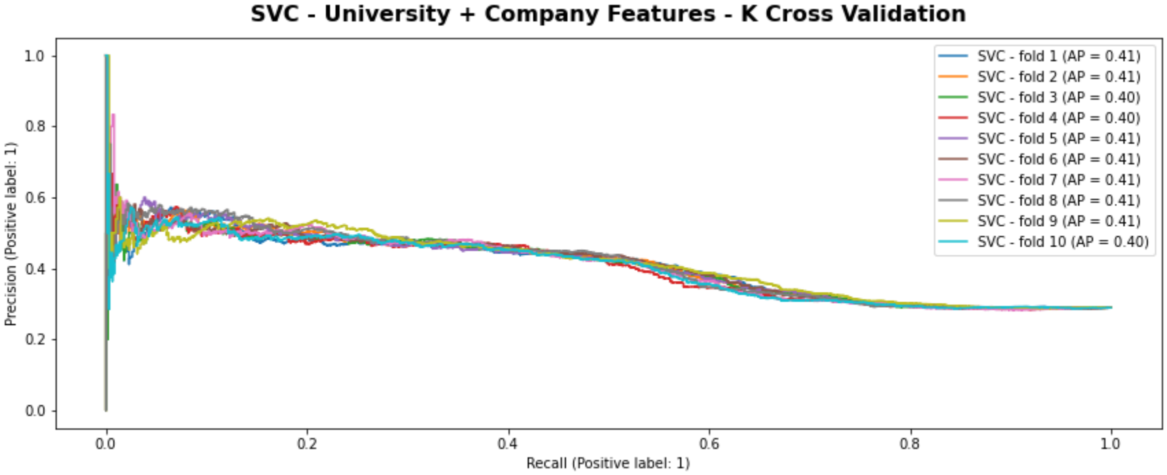
$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)}$$

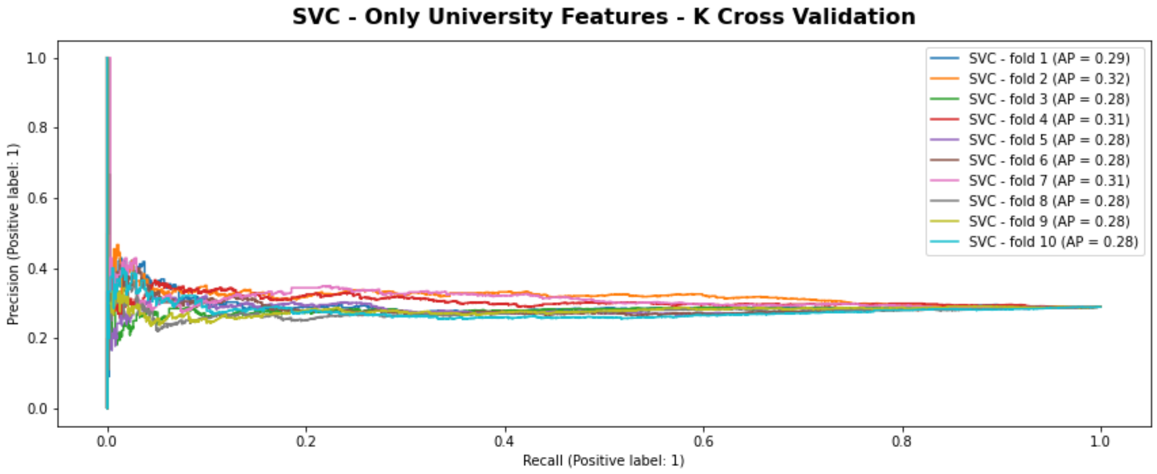
$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)}$$

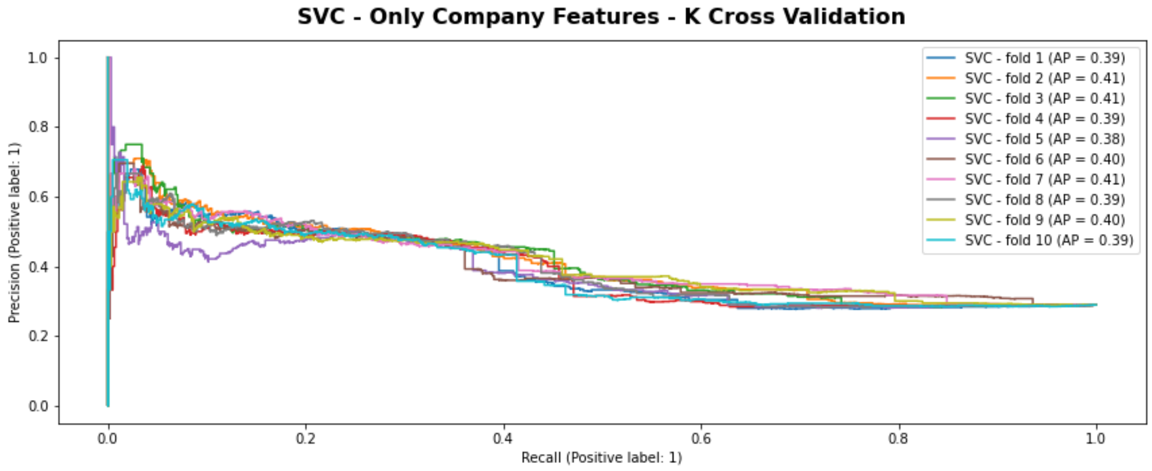
A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will return many results, with all results labeled correctly (sklearn). Therefore we are most interested in generating models with both high precision and recall, but precision should take priority.

4.2 Support Vector Machine (SVC)

First we standardise all parameters so they're values between 0-1. We group the selected features into two clusters, University and Company, and run SVC model on them separately and combined.



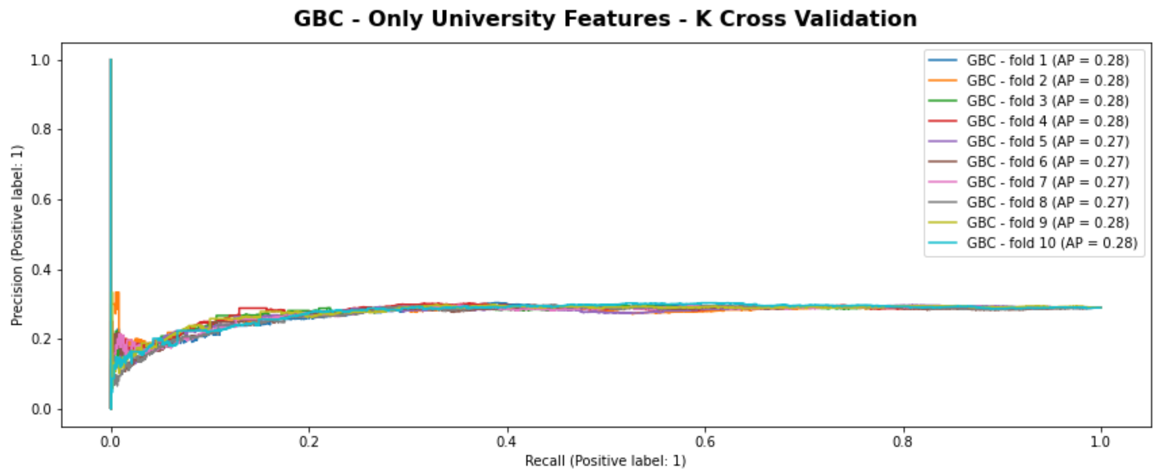


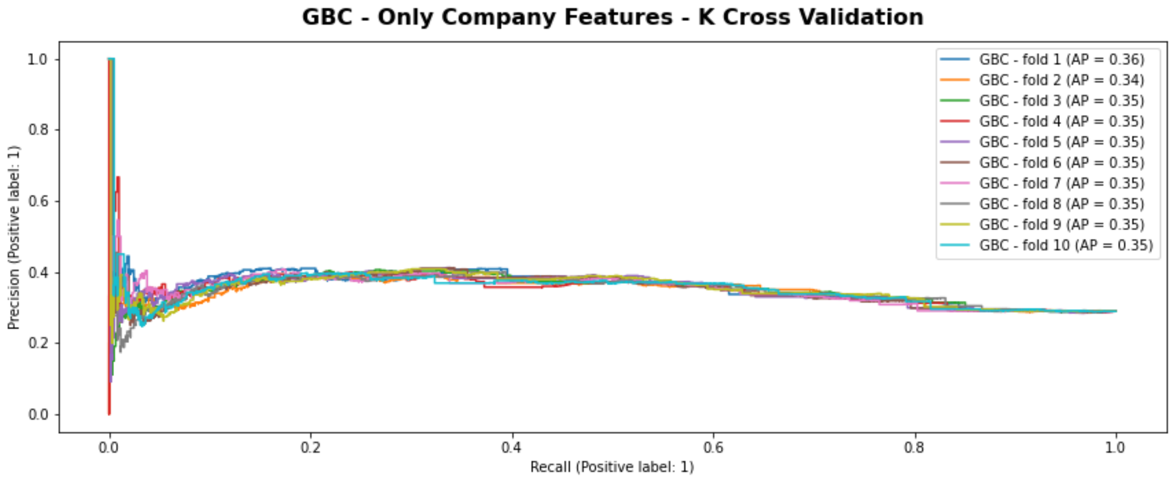


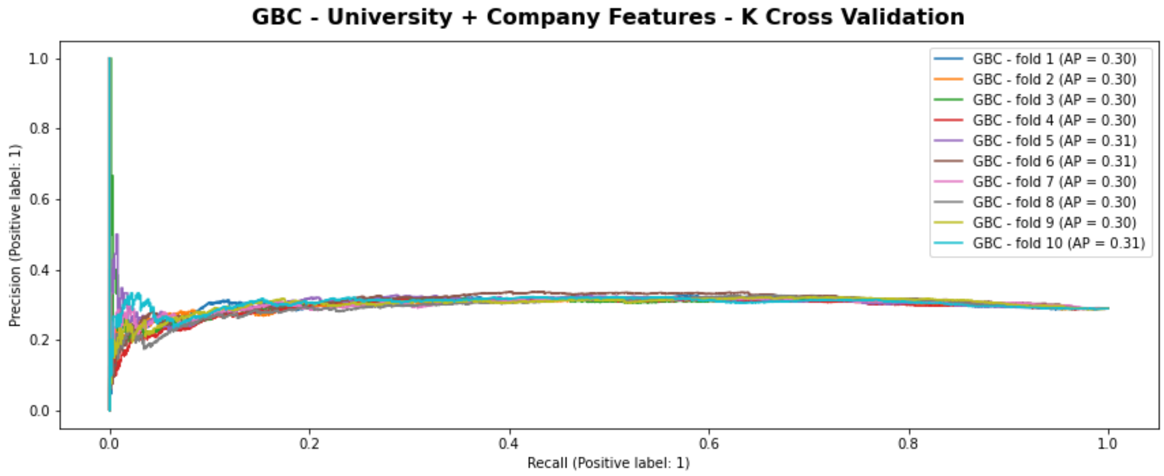
From figures, we see good combinations of high Precision and Recall (0.4 Precision at 0.4 Recall) from both the Company Features Only trial and the combined trial, while University Features Only have worse Precision even at lower Recall rates.

4.3 Gradient Boosting Classifier (GBC)

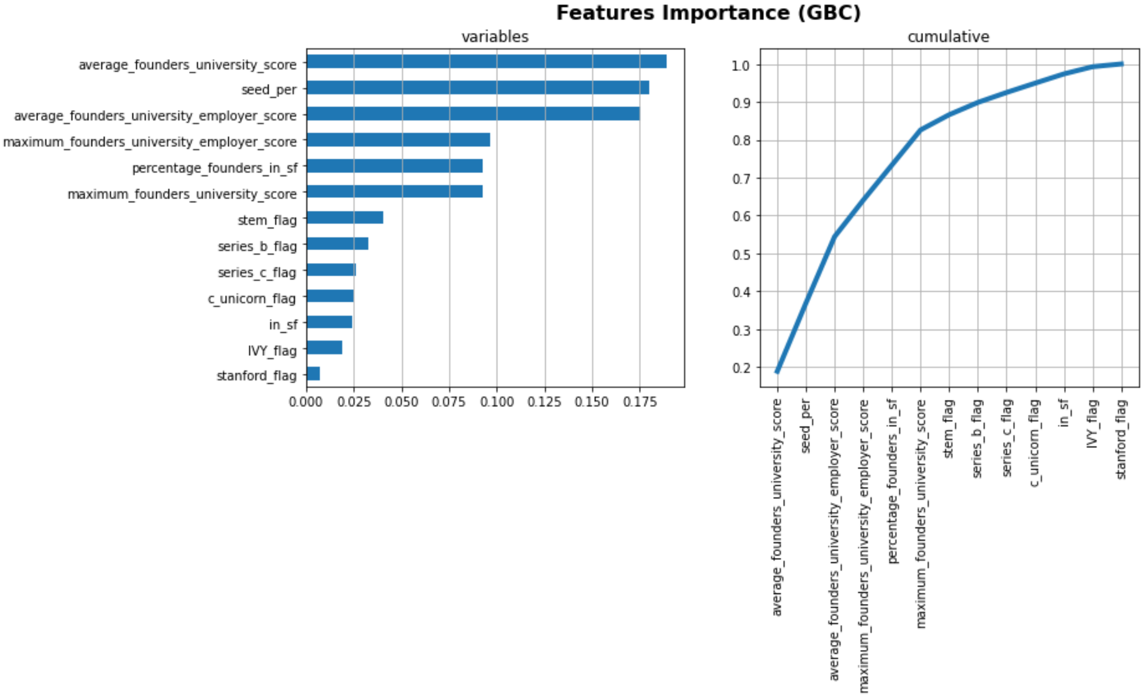
Similarly, we run Gradient Boosting Classifier (GBC) models on the standardised data and 3 trials of University Features, Company Features, and Combined Features.







From the Feature Importance Rankings for the GBC model, we see that University Score, whether founders had previous experience in seed stage startups, and founder's University Employer Reputation are all parameters that differentiate whether a company is invested by a successful or failed investor. It is interesting to note however, that features like based in San Francisco, or having founders from San Francisco, don't have high importance regardless of their high correlation to 'success_flag' as discussed in Section 3.2.



5 Conclusion

In this report we try to find features of start ups that distinguish between the proxy of whether it is invested by a successful or brand or failed investor. Through statistical analysis, feature importance ranking and feature engineering, we establish that there are 3 major areas of features that contribute to success correlation:

- Previous experience in start-ups, where founders' previous experience in start-ups that became Unicorns or reached Series C and B funding has a very high correlation to their current company being invested by successful investors; while founders that previously worked in seed round start-ups (i.e. didn't progress further) has a very high negative correlation to being picked by successful investors.

- University: features like University ranking, particular Universities (Stanford, IVY league), University Employer Reputation, and inclusion of a STEM subject all have positive correlation to success. Features like highest degree obtained and International Universities gave no correlation.

- Geography: Companies in San Francisco tend to be picked by successful investors, this trend does not extend to other major cities (New York, Miami). However comparing datasets from before and after Covid shows this advantage is diminishing.

Running SVC models using selected feature clusters (University, Company) shows that both Company and combined features yielded more accurate results.

Running GBC models, we see a similar distribution of feature importance, one thing to note is being in San Francisco is no longer an important feature, despite the feature having high correlation to success. Future research can be conducted to investigate the difference in behaviour between the ML models and traditional statistical analysis.