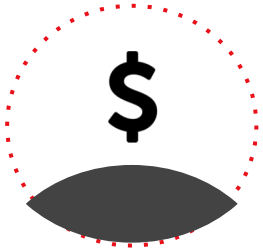
Startup: Success or Failure?

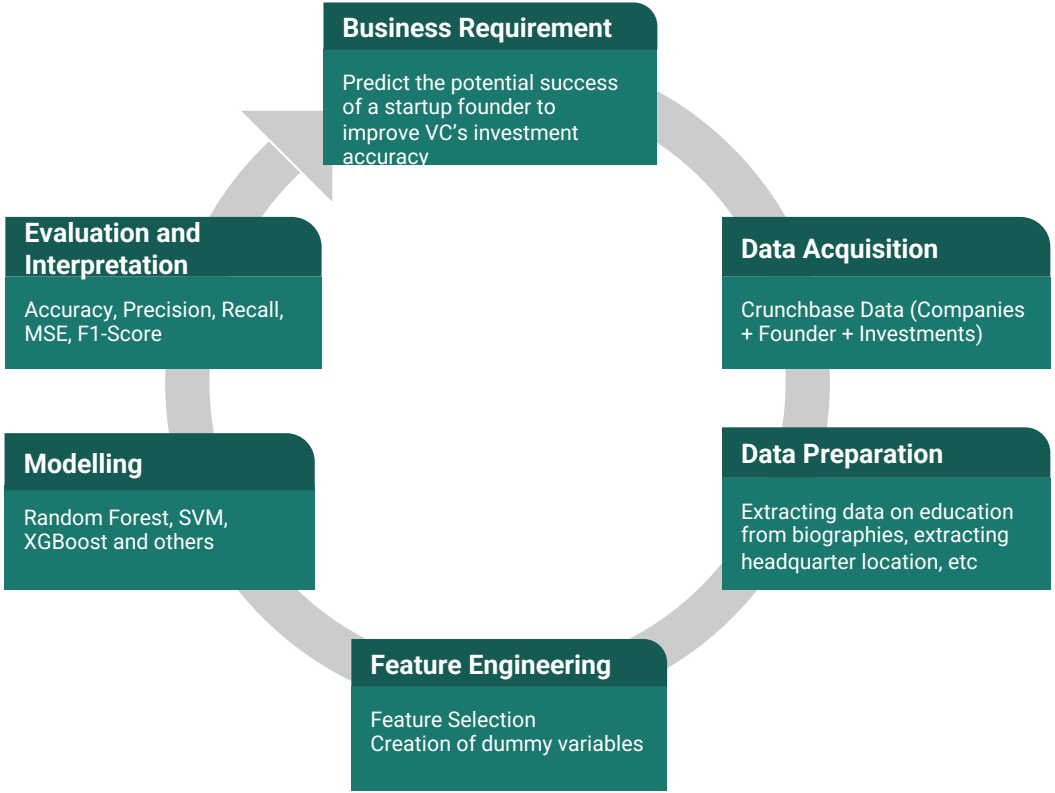
Value Proposition



Helping Venture Capitalists improve accuracy of their investments by understanding whether a start up will be successful

(Raising minimum Series B funding)

Our Approach



Data Extraction

Data was extracted from Crunch base and GitHub

There were 3 primary data files that had information on:

1. Companies (Industry, Headquarters, Founded Date, Company Type)
2. Founders (Biography, Gender, Number of Founded Organizations, Number of Portfolio Companies)
3. Investments (Funding Type, Funding Date, Funding Amount)

Data Manipulation

Extraction of Data on Education

- Extract education by merging on Founder Name from Crunchbase website.
- Extract data on whether founder attended ivy league college using bio column
  For example, *'Sergey Sundukovskiy, Ph.D. has over 20 years of experience serving*…' would give a 1 value for Ph.D.

Feature Engineering

- **Dummy on Masters, PhD**

  Bachelors is the base case

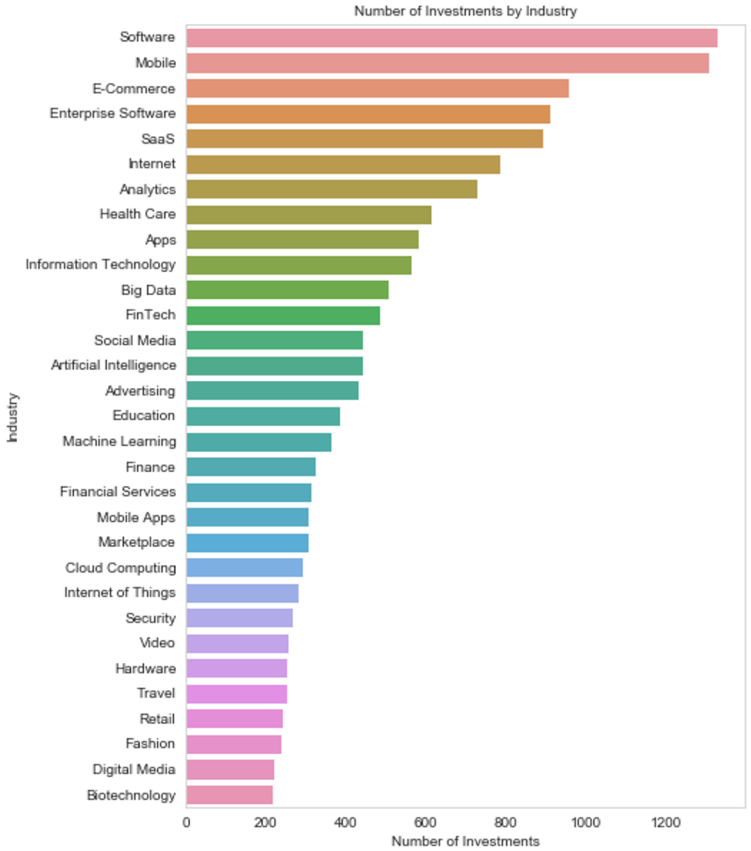
- **Creation of headquarters feature**

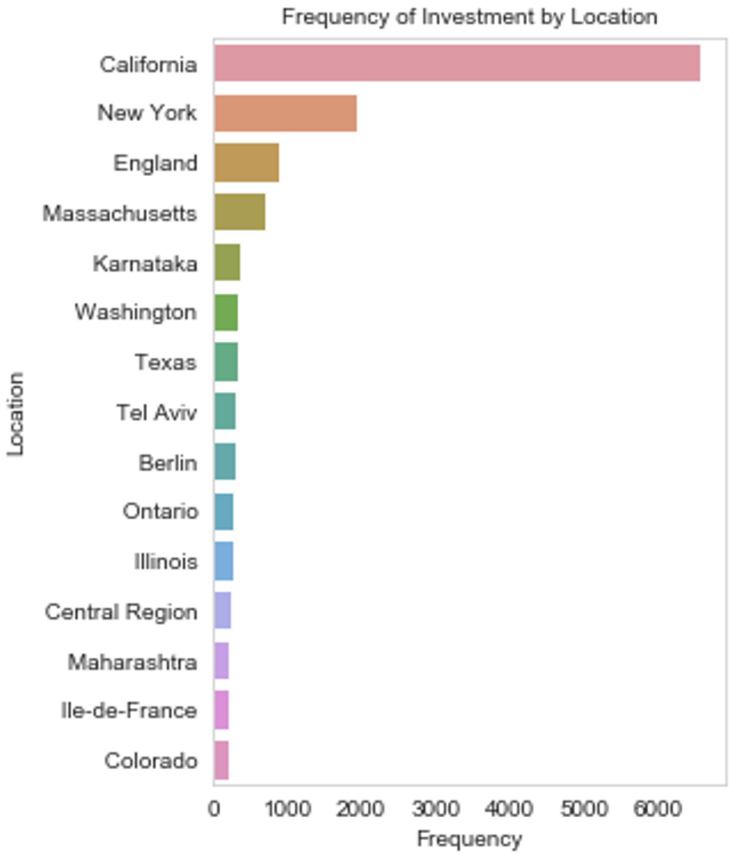
  We replaced the headquarter column which was in the format of *'City, State, Country'* with the frequency of occurrence of that location.

- **Dummy on Ivy League + Stanford + MIT + Caltech**
- **Dummy on industry**

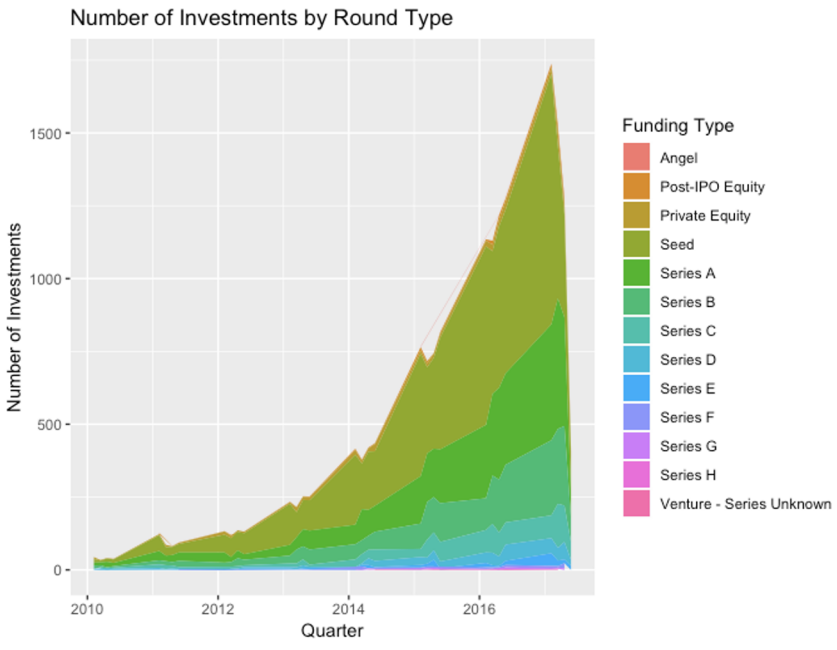
  The main 11 categories are Tech, Finance, Healthcare, Logistics, Real Estate, Retail, Lifestyle and Fashion, Travel and Hospitality, Energy/Environment, Media and Marketing and Other

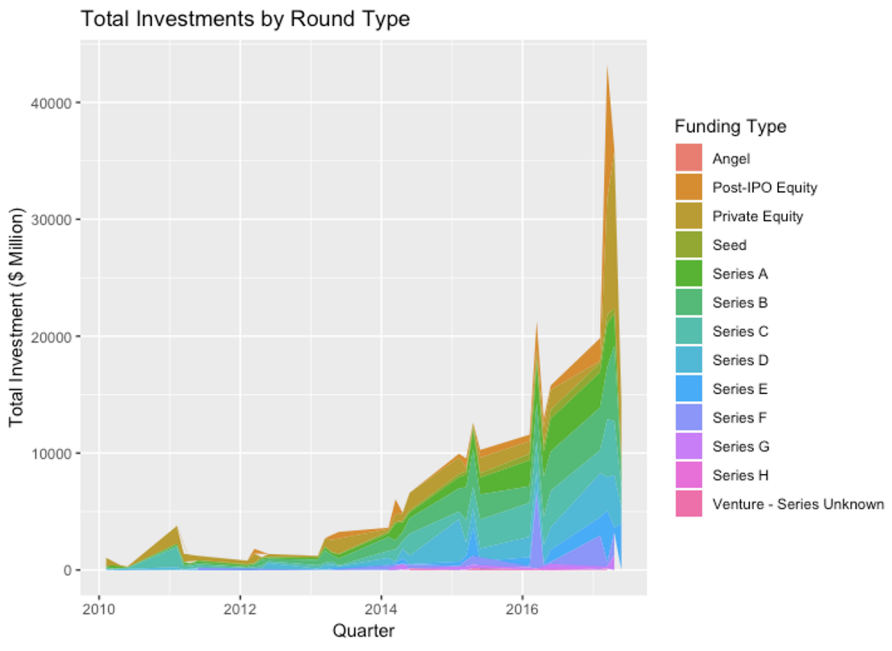
Data Visualisation



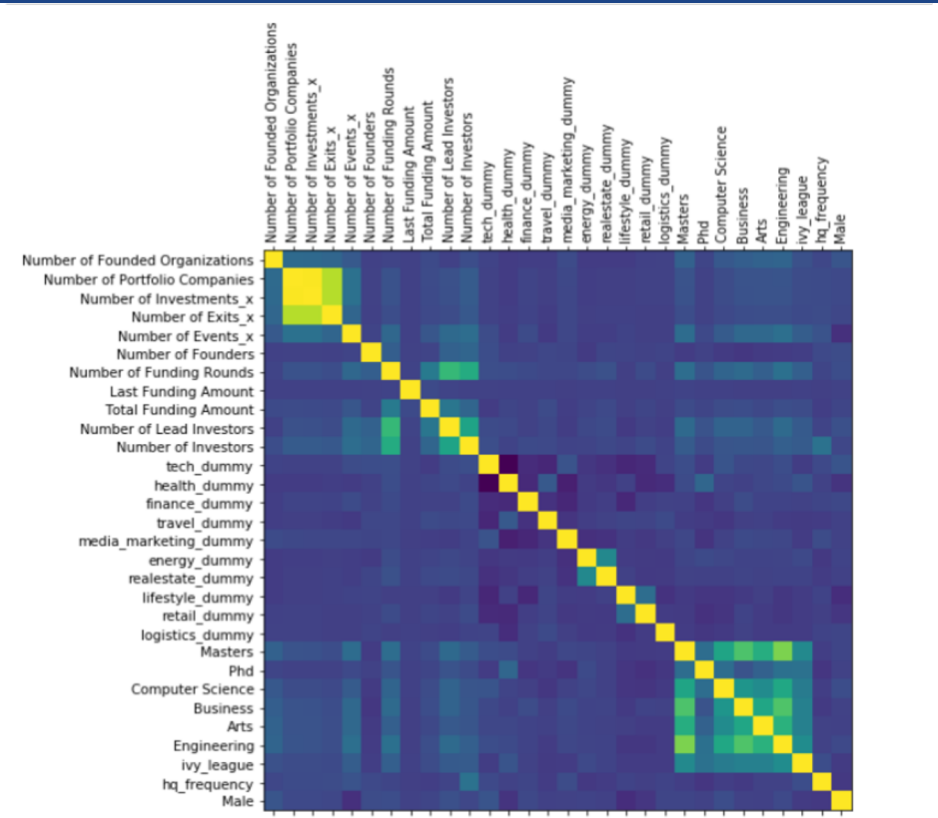


Investments by Round





Correlation Matrix

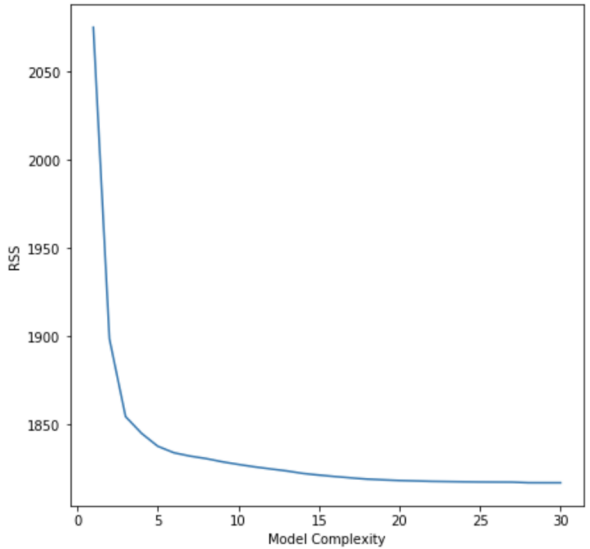


After analyzing the correlation matrix, we dropped highly correlated features such as:

1. Number of lead investments
2. Number of partner investments
3. Last equity funding amount
4. Total equity funding amount
5. other dummy
6. bachelor dummy

Forward Stepwise Selection

Forward stepwise selection: We ran forward stepwise selection on all the features, and plotted the minimum RSS of each model containing (1… 30) number of features. We chose the n that minimised the RSS. The final number of features are and 8 features chosen are:

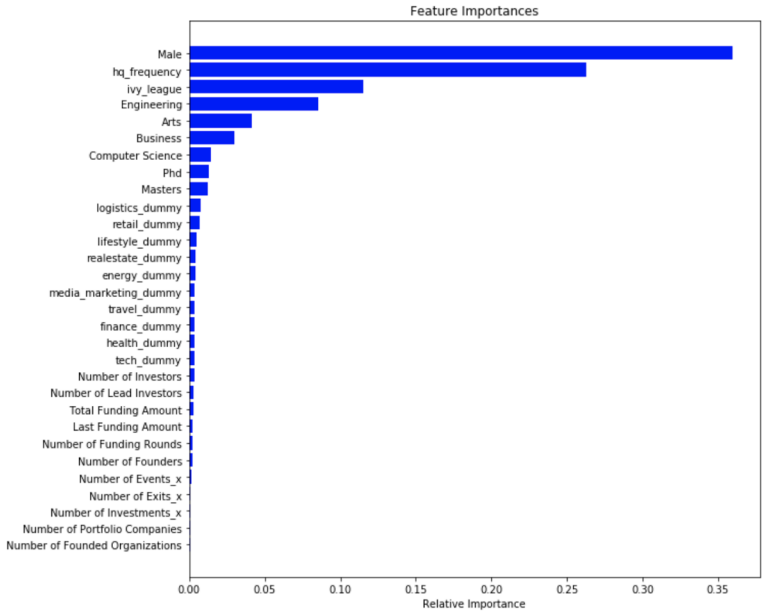



```
-----------------
Foward Selection:
-----------------
Number of Lead Investors
Number of Funding Rounds
Masters
hq_frequency
ivy_league
Computer Science
Male
Number of Exits_x
Number of Founded Organizations
Business
```

Random Forest

We divided the model into 75-25% train test and ran random forest on it, using all the features in the dataset except target variable. We performed a Grid Search Cross Validation on the model, and got the the best score and the best parameters.

```
The best parameters are:  {'max_depth': 15, 'min_samples_leaf': 4,
 'min_samples_split': 8, 'n_estimators': 50}
```



XGBoost

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

The best scores and parameters obtained through this model are as follows

```
Model with rank: 1
Mean validation score: 0.919 (std: 0.004)
Parameters: {'colsample_bytree': 0.9063511098987536, 'gamma': 0.4738461983918303, 'learning_rate': 0.296172775841290
54, 'max_depth': 5, 'n_estimators': 143, 'subsample': 0.6781136557857271}
```

SVC

A Support Vector Classifier (SVC) is a discriminative classifier formally defined by a separating hyperplane. Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

We used the cross validation grid search to find out the hyper parameters. The best kernel coefficient is 0.1 and best penalty parameter C is 10.

```
1  svc_search.best_score_, svc_search.best_params_
```

```
(0.8313583815028902, {'C': 10, 'gamma': 0.1})
```

The accuracy of our SVC on training set is 84.96%

Logistic Regression

The coefficients for Logistic Regression are as shown

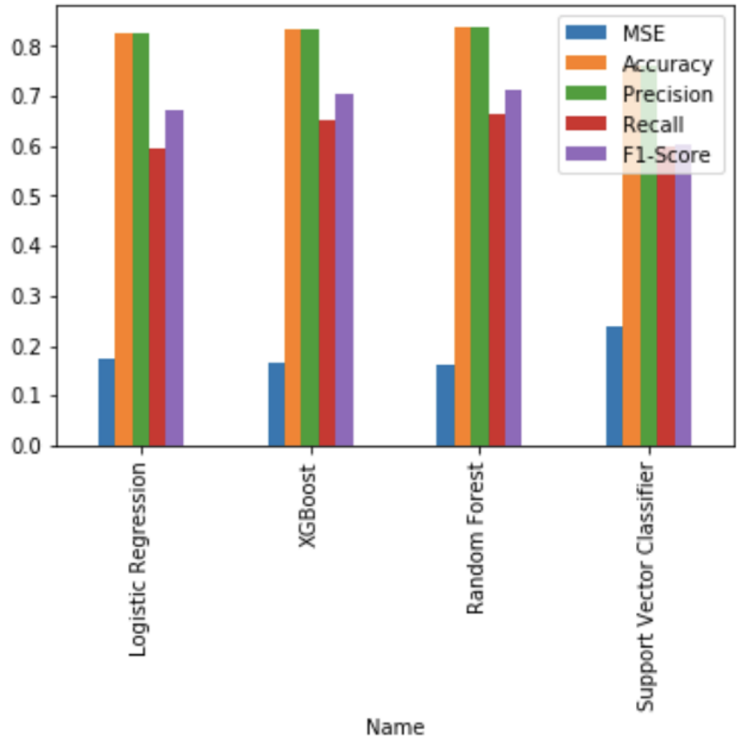
```
{'Arts': -0.044750991111363735,
 'Business': 0.193135911777947,
 'Computer Science': 0.22338590451682291,
 'Engineering': 0.1414911000626613,
 'Last Funding Amount': 0.47074045579130763,
 'Male': 0.28353125444912664,
 'Masters': 0.30308613441550253,
 'Number of Events_x': 1.2322971346269047,
 'Number of Exits_x': 1.1712374520918065,
 'Number of Founded Organizations': -1.0100408934357101,
 'Number of Founders': -0.26392842652377024,
 'Number of Funding Rounds': 9.748453889688799,
 'Number of Investments_x': 1.1299853714648207,
 'Number of Investors': -0.3119047183547719,
 'Number of Lead Investors': 12.814554101442019,
 'Number of Portfolio Companies': 1.282213141367116,
 'Phd': 0.20004151462538092,
 'Total Funding Amount': 3.8085185457044757,
 'energy_dummy': -0.06512719840479281,
 'finance_dummy': -0.22554820048061272,
 'health_dummy': -0.06084326547791157,
 'hq_frequency': -0.5924665568756124,
 'ivy_league': 0.3789985429342101,
 'lifestyle_dummy': -0.16866793423422588,
 'logistics_dummy': -0.13464686360957157,
 'media_marketing_dummy': 0.15813905725784352,
 'realestate_dummy': 0.032750652414070866,
 'retail_dummy': 0.19024572811330098,
 'tech_dummy': -0.028289165947819717,
 'travel_dummy': -0.0817171081792974}
```

Business, Computer Science and Engineering have a positive coefficient whereas Arts has a negative coefficient showing founders having CS, Eng, Business degrees succeed

All the results are intuitive, except for tech_dummy, number of investors and Headquarter frequency which have a negative coefficient.

Performance from K fold cross validation

	Name	MSE	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.172907	0.827093	0.827093	0.597859	0.676666
1	XGBoost	0.075755	0.924245	0.924245	0.857003	0.872584
2	Random Forest	0.085875	0.914125	0.914125	0.822116	0.852860
3	Support Vector Classifier	0.169293	0.830707	0.830707	0.617855	0.688341



Model Analysis

Accuracy of the baseline model is 82.44% and MSE is 17.56%

Best choice: XGBoost and Random Forest

```
Accuracy of Test Set using XGBoost is 92.65%
MSE of Test Set using XGBoost is 7.35%
```

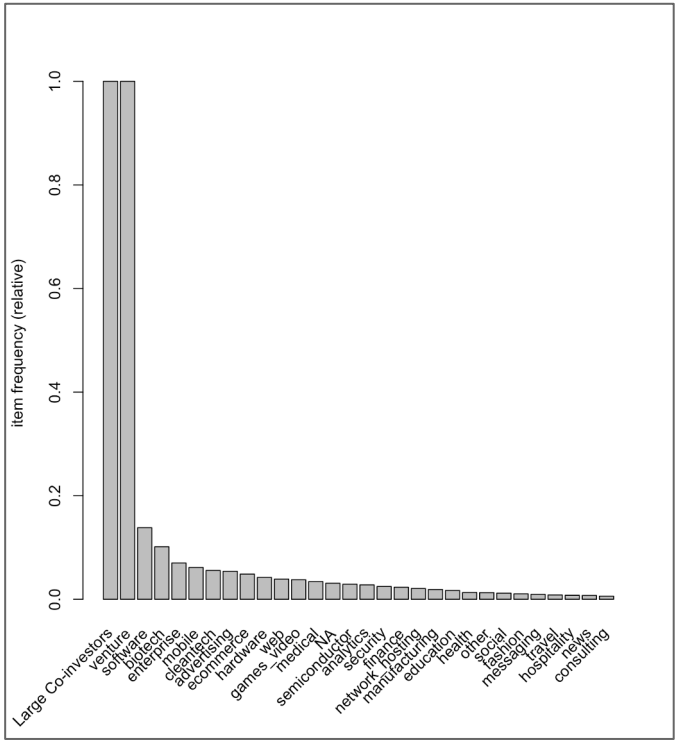
```
Accuracy of Test Set using Random Forest is 90.89%
MSE of Test Set using Random Forest is 8.85%
```

While Random Forest is easier to run and gives feature importance, XGBoost is more accurate

Market Basket Analysis

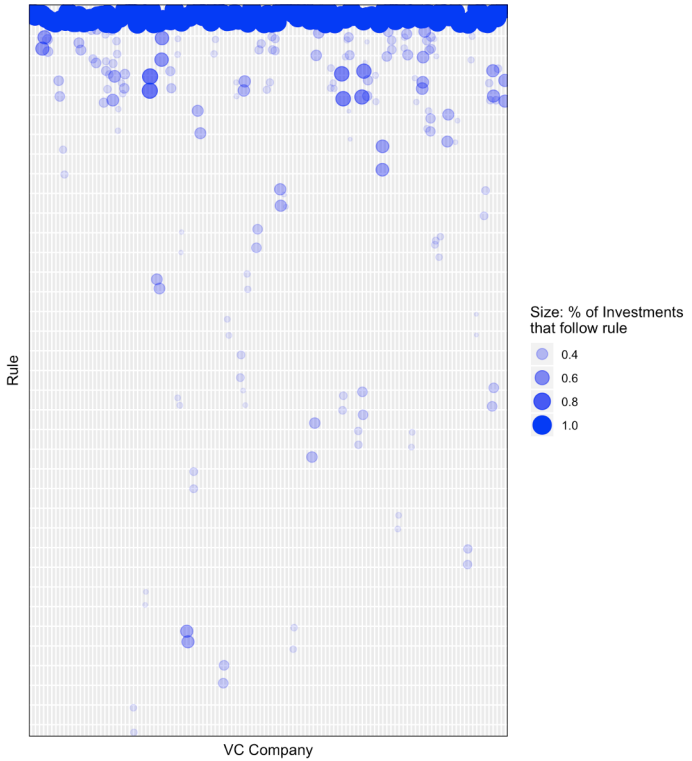
An algorithm to detect association rules (which features occurred together for a successful startup) and get feature likelihood, to understand to what extent the factors occur simultaneously for a VC portfolio.

Some investing features are quite common across VC portfolios. For instance, Large Co-investors are present in more than 80% of portfolios.



Rule usage per VC investor

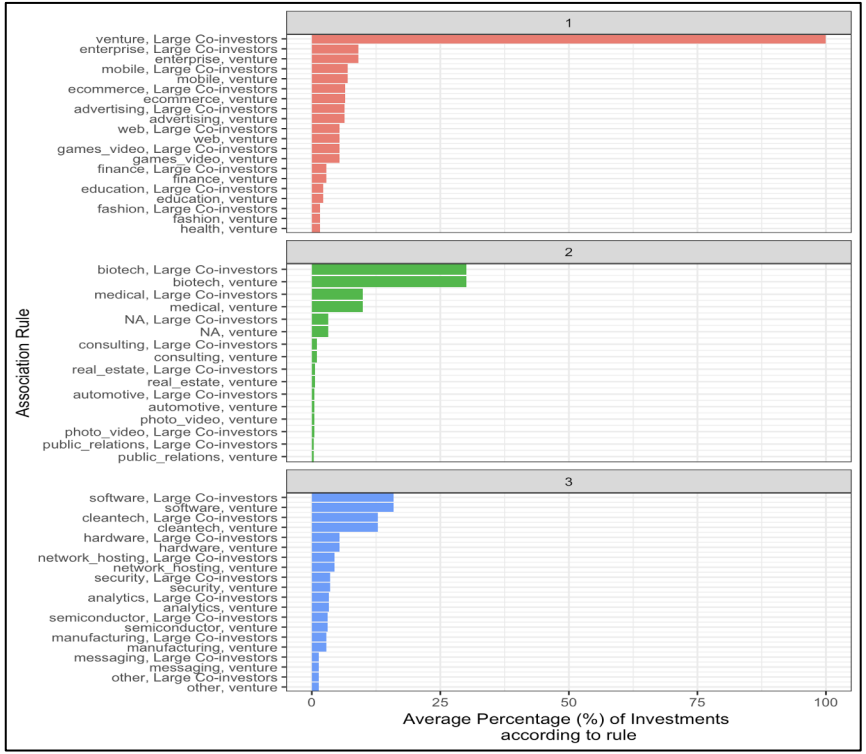
The vertical axis contains each rule, whereas the horizontal axis contains each VC firm in the dataset. The circles' width represents the percentage of the company's portfolio that followed the corresponding rule



Hierarchical Clustering

Uniqueness of Clusters

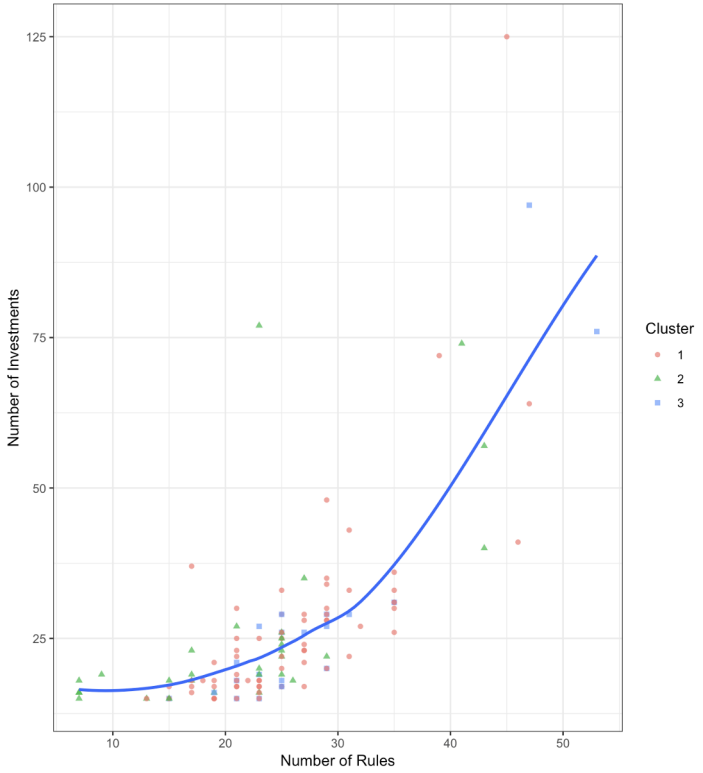
For each rule which clusters have the highest average percentage of use (used two-sample test of proportions to filter only the rules that are significantly more preponderant in each cluster)



Portfolio Diversification

The largest VC firms in terms of the number of investments used the highest number of different rules which suggests portfolio diversification

Most of the largest VC firms were classified in cluster 3, which provides insight into the role of these firms in financing late stage rounds and software companies.



Insights

1. Even though the number of founder with business degrees is higher, we were able to deduce that the correlation between a founder succeeding if he has a computer science background is higher with good confidence. This can be due to the growing digital/software/data related startups

1. There is a good chance of a founder succeeding if his alma mater is an Ivy-League/Top-School contrary to the popular belief of being a dropout. We hypothesize this can be true given the great alum network or maybe just the nature of the network a founder is exposed to in an Ivy league

1. Number of founded organisations is negatively correlated which can be attributed to distraction

1. More number of founders is negatively correlated to success. This is consistent with the idea: too many views (people) result into poor decisions

1. Startups with highest number of investments are focussed at portfolio diversification as observed in our market basket analysis

1. Most of the largest VC firms finance late stage rounds and tech companies. So more and more tech startups are blooming and most new start-ups are in the tech category

Thank You