ISM 6136 DATA MINING FINAL PROJECT START UP DATA ANALYSIS AND PREDICTIVE MODELLING

Suryanarayana Aneesh Prayaga(U93559651)

Amulya Alwala(U39022318)

Raj phanindra Kakarla(U04426707)

Background of estimating the startup success:

In modern era, startups are the engines of growth and innovation. They are the driving force of economy. Though startups are small they lead to major transformations in terms of technology, way of living and employment generation. With increasing number of entrepreneurs making their startup dream a reality to quench the need of the fast-pacing world, it has become increasingly important to support and enable the success of startups. This starts with estimating their probable success or failure, to help them come up with better strategies to sustain and conquer the market. Also, helping the investors to make smart investments to mitigate risk.

Motivation behind solving the problem:

The motivation behind solving this problem is there are a numerous of startups which are failing due to incorrect analytics and setting up in a wrong geographic location. Through this ML model we want to analyze the existing data and formulate a model which could predict if a startup is going to be successful. Our intention is to provide a good and practical starting point for a upcoming startup to weigh their options, to open up in a given location or foresee the risks that could occur in future and keep them prepared. There are mainly 3 advantages to formulate this model.

1. Before launching a startup, what could be the factors an entrepreneur needs to consider this can be based on the company's parameters (Location, Industry, Funding, Relationships etc.)
2. We would also like to provide a startup "Checklist" for a company based on Demographic or category or any other factors in our data set and increase its odds of success and help the company set a long-term goal.
3. If a company is in the market already and if the company wished to evaluate its position and forecast its possible future this model can provide a comprehensive idea.
4. It can also showcase the types of industries based on parameters and cluster them to get general idea on the trends in the industry.

This analysis could be crucial because there are innumerous startups taking birth every month, but a very few will make it to the market and survive. Our effort and ultimate goal is to increase the success percentage of startups by proving a blue print for success.

Solution Methodology and Evaluation Metrics:

In context of estimating the probable success or failure of a startup, we have considered the status variable which has acquired and closed values. The former signifies the success of a startup from financial gain through merger and acquisition and the latter signifies the eventual shutdown and failure of the startup.

To achieve this, we have performed extensive exploratory analysis using K-means clustering on the data set consisting of more than 40 variables using R as well as Azure ML studio. Following are some high-level outcomes of clustering

- Exploratory analysis to find the trend of different variables and their possible impact in the outcome of a startup's success or failure
- Identifying the probable relationship between different variables or their impact on the outcome, that couldn't be discovered otherwise.

After exploring and understanding the dataset and comparing different algorithms, we came up with the following major classifiers that gave better results.

- Two Class Decision Forest
- Two Class Boosted Decision Tree

As our expected outcome is to classify the start up in terms of two statuses, acquired and closed, we found two class classifiers to be ideal to estimate the success using these values.

Clustering:

The Clustering model is Built on K means Clustering, to derive the Optimal number of clusters in the Dataset We have used a Scree plot and from this we have concluded that the number of clusters could be 3 or 4.

Methodology of clustering:

1. Plotted a scree plot to get the least sum of squares (WCSS) .
2. Change the categorical variables to binary using Categorical variable encoding.
3. Set up K means clustering and visualize the Data.
4. Use PCA to explain the variation between data and find which variable/feature contribute most to the clustering or explains the clustering better.
5. Derive insights from the Clustered Data and explore.

Data Set Description:

The dataset consists of mainly consists of 9 Features and in which Status will be the success variable.

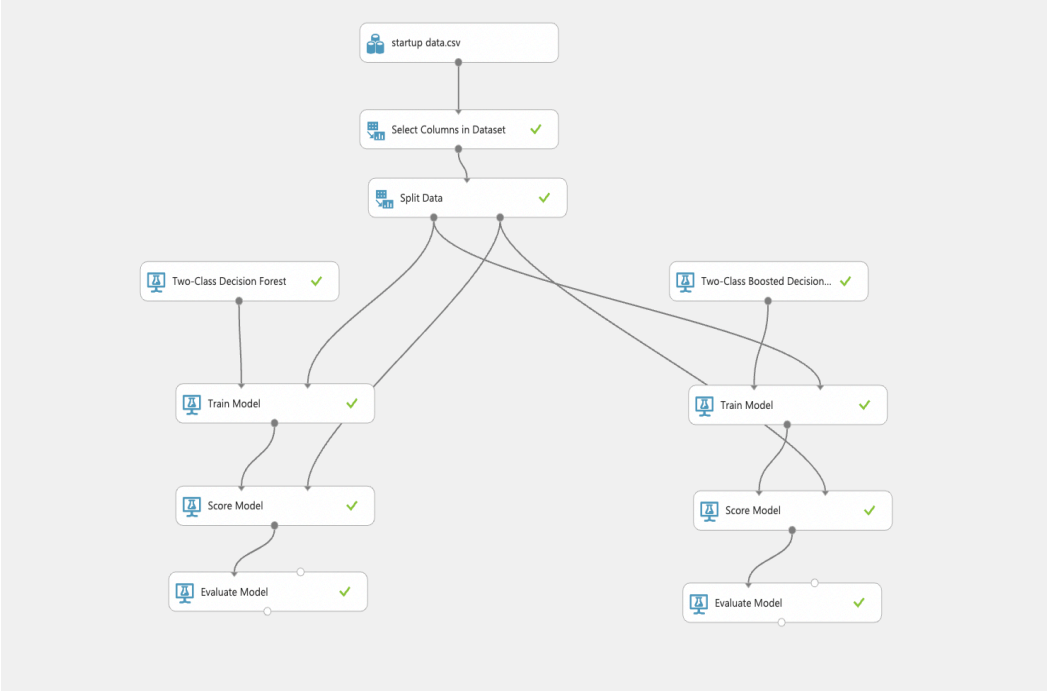
State code (Categorical): The state code consists of number of states in the data set. In the dataset we have mainly focused on CA, NY, MA, TX and other states.
Relationships (Numerical): This is a numerical variable which provides number of relations an existing company holds with other companies.
Funding rounds (Numerical): The number of financial rounds for the startup.
Category code (Categorical): The area of business the company is operating in. In this Dataset we have focused on Advertising, Mobile, enterprise, Software, Web)
has_VC: this column shows if the company has any venture capitalists involved.
Has_angel: This column shows if the company has angel investors.
Average participants: This column shows the number of co-founders or principle working group of the company.
Is_top500: This column shows if the company made it into top 500.
Status: This column shows the success or failure of the startup. (Acquired or not acquired by a company).

The data set is made sure that the within correlation of the features is minimized and the correlation with the output variable to the features is minimized.

Comparison of Algorithms:

The analysis of two class decision forest and two class boosted decision tree was based on terms of different parameters like, Area under the curve, accuracy, precision, the number of true positives, false positives, true negatives, and false negatives.

The screenshot below shows the implementation of these two algorithms using different parameters



TWO CLASS DECISION FOREST:

Following are the parameters selected and their corresponding outcomes for two class decision trees

Since the dataset has many startups that are acquired and successful and prioritizing the correctly predicting the probability of success, and eliminate the incorrect estimation of success, we are inclined towards a model with a greater number of true positives, false positives, better AUC, and accuracy.

Properties Project

Two-Class Decision Forest

Resampling method: Bagging

Create trainer mode: Single Parameter

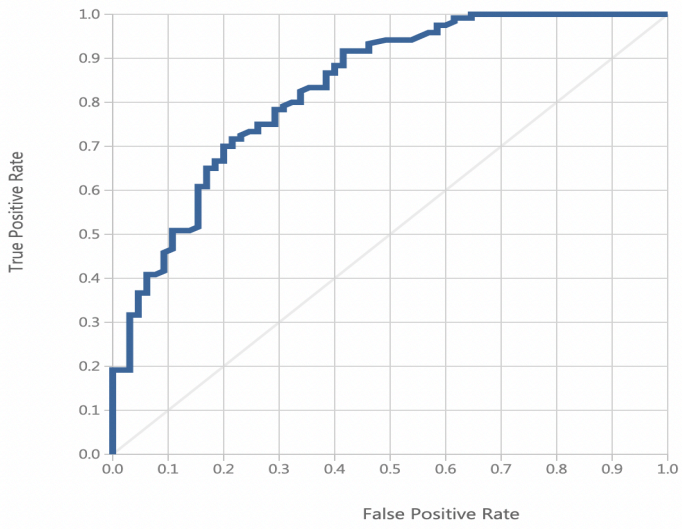
Number of decision trees: 8

Maximum depth of the de...: 16

Number of random splits ...: 128

Minimum number of sam...: 1

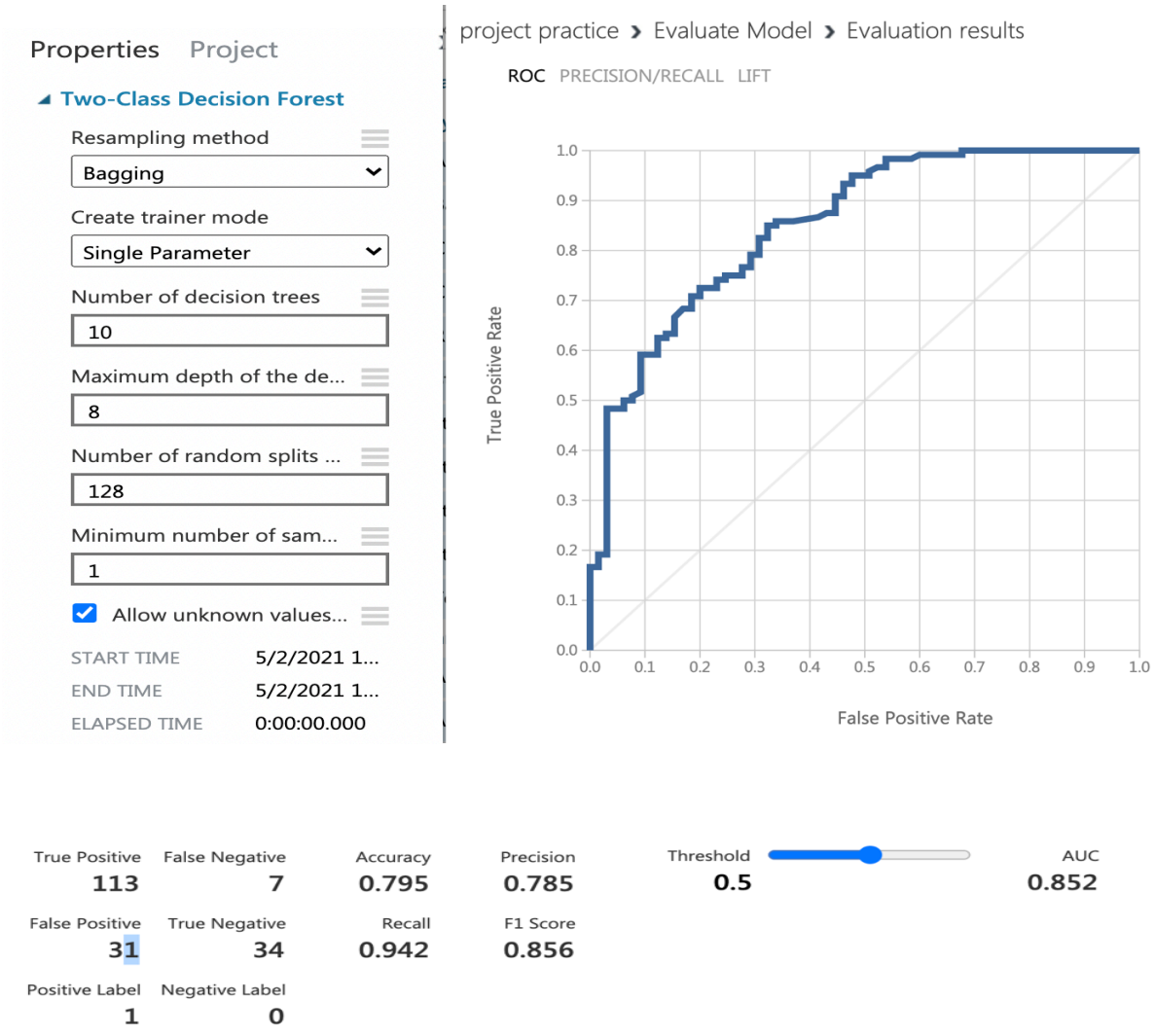
ROC PRECISION/RECALL LIFT



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
110	10	0.784	0.786	0.5	0.831

False Positive	True Negative	Recall	F1 Score
30	35	0.917	0.846

Positive Label	Negative Label
1	0



Model 1:

This model is parameterized with 8 decision trees and a maximum depth of 16, as per the output below, we can notice the AUC is 0.83 with 110 true positives.

Model 2:

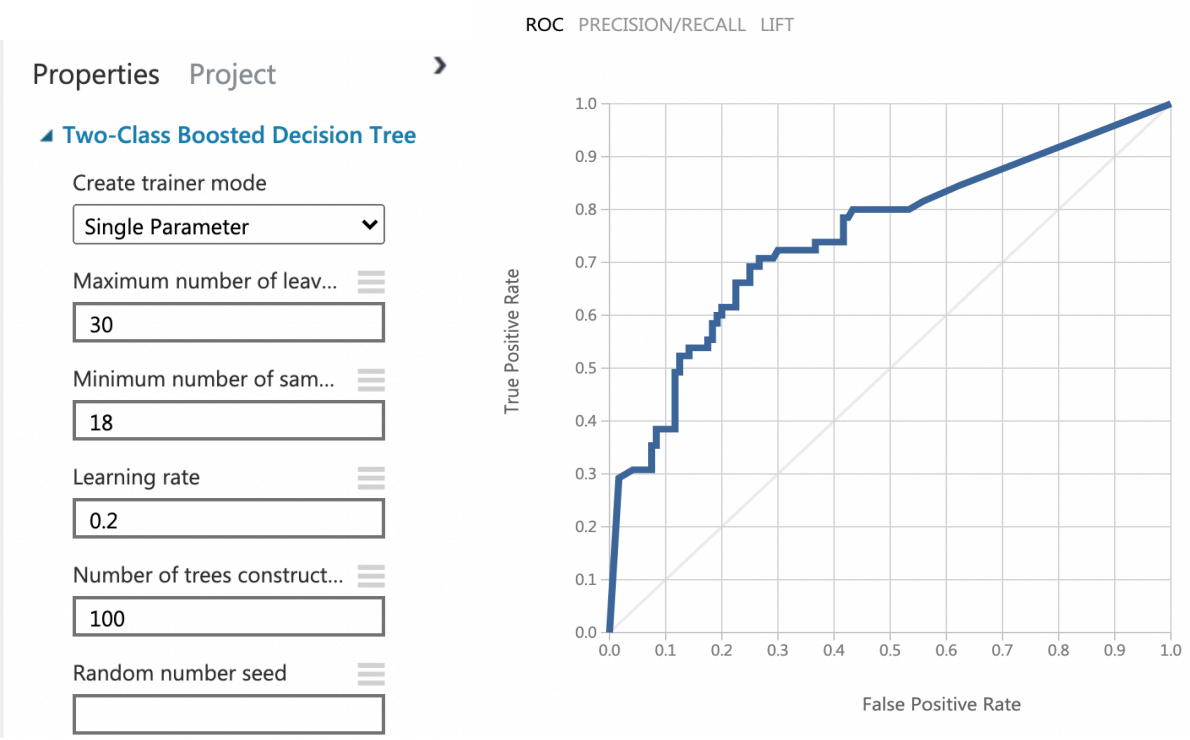
This model is parameterized with 10 decision trees and a maximum depth of 8, as per the output below, we can notice the AUC is 0.852 with 113 true positives.

In context of the above-mentioned outputs, model 2 is more ideal fit for our data set and to achieve the required outcome for our problem, owing to the high area under the curve and greater number of true positives

TWO CLASS BOOSTED DECISION TREE:

Following are the different parameters selected for two class boosted decision tree and corresponding results

Model 1:

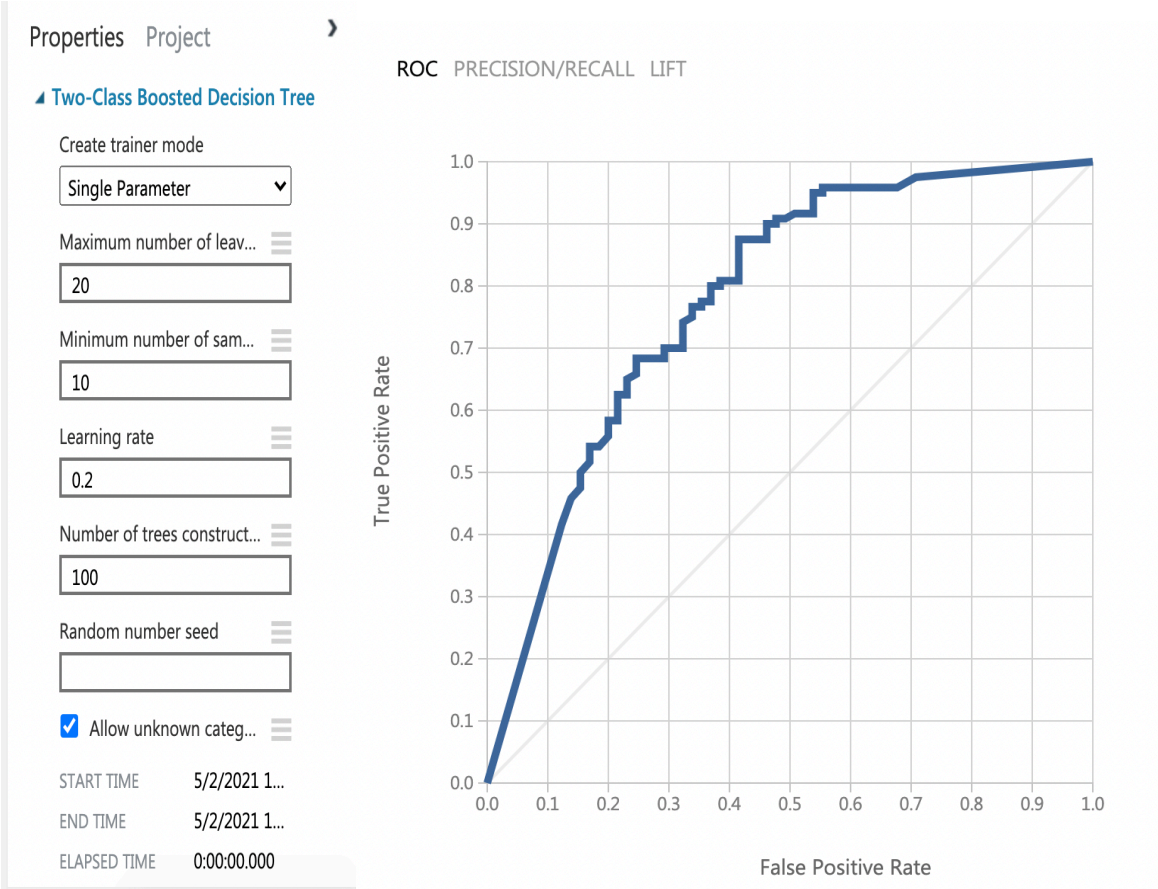


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
39	26	0.730	0.619	0.5	0.751

False Positive	True Negative	Recall	F1 Score
24	96	0.600	0.609

Positive Label	Negative Label
closed	acquired

Model 2:



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
101	19	0.751	0.789	0.5	0.769
False Positive	**True Negative**	**Recall**	**F1 Score**		
27	38	0.842	0.815		
Positive Label	**Negative Label**				
1	0				

Model 1:

This model is parameterized with 30 maximum number of leaves and 18 minimum number of samples, above output shows the AUC is 0.751 with 39 true positives and false positives 24.

Model 2:

This model is parameterized with 20 maximum number of leaves and 10 minimum number of samples, above output shows the AUC is 0.769 with 101 true positives and false positives 27.

As per the above-mentioned values, model 2 is better suitable to achieve the required outcome for our problem as it has high area under the curve and greater number of true positives and false positives

Comparison of Two class decision forest and Two class boosted decision tree:

Model 2 of two class decision forest is more ideal and viable for our problem as it has greater accuracy, AUC, and far greater values of True and false positives

Two Class decision forest analysis:

Since the two-class decision forest generates multiple decision trees as per the parameterization with multiple starting points and the entire data set and performs voting for classification of a new value as part of bootstrap aggregation. Following is the classification and analysis based on one such decision tree out of the 8 for the trained model.

The model initial chooses the starting point with number of relationships less than or equal to seven.

The tree is then split in terms of average participants for values with relationship less than 7 and category code security.

Considering the relationships less than 7, it is further filtered based on category code hosting, upon satisfying this condition it is assessed based on is_top500 variable, which indicates if the company is ranked within top 500 or not.

If the company falls under top_500 category, the model progresses with the next condition of checking for the industry type photo video using the variable category code.

Finally, the model predicts that the startup company in photo video industry can be acquired, indicating good chances of success such startup company which falls under the above category and conditions.

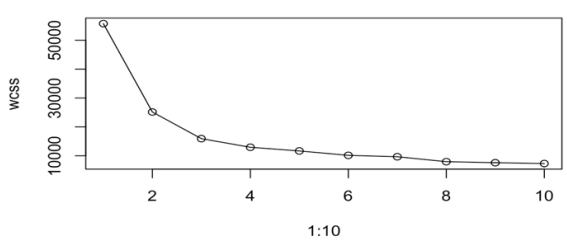
Similarly, the model has predicted if the status of the company will be closed, indicating the probable failure of the company.

Clustering:

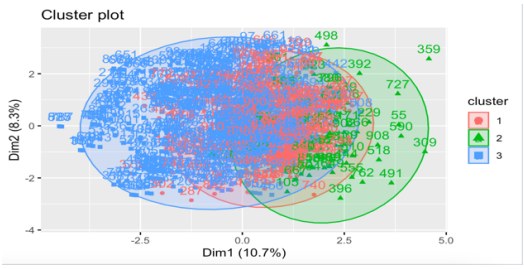
The clustering model can also be used to give suggestions on the factors that play a pivotal role in determining the potential success of a company or which factors the competitors have their strength in.

Outputs of Clustering:

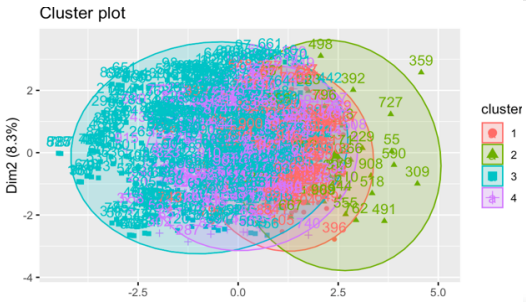
Scree Plot



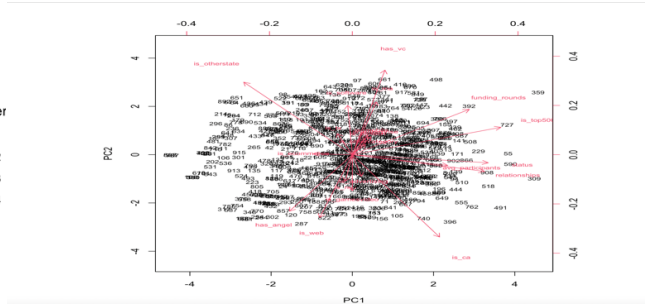
Data With 3 Cluster



Data with 4 Clusters



PCA Plot

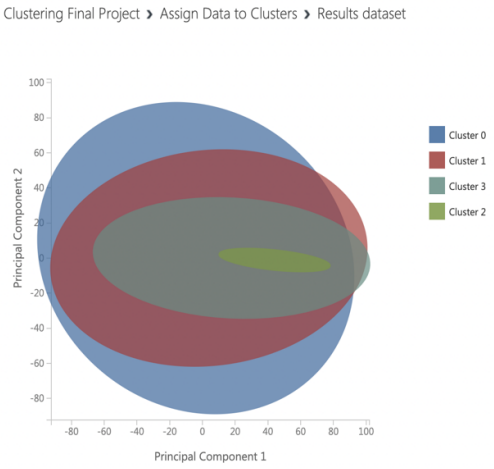


PCA Analysis: From the PCA analysis the below 4 variables account for the maximum variation along the Dataset
1)Is CA
2)has_VC,
3)Relationships
4)Funding Rounds.

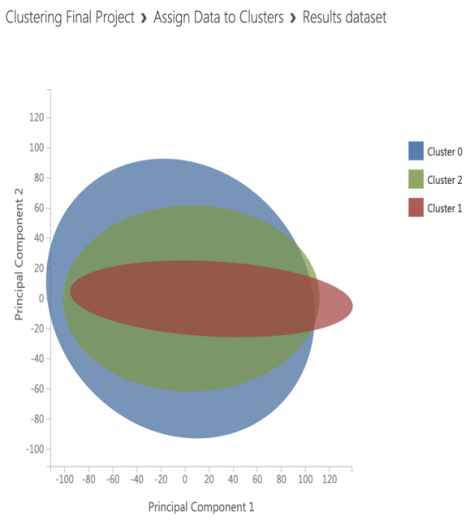
If, we were to do further Analysis and more cleaning and analysis of data, We could derive insights like which parameter is causing the variation ,which parameter contributes the least and the general trend of the industry and data.

Clustering Outputs in Azure :

K=4 Clustering



K = 3 Clustering



Analysis from Clustering:

1. The number of relationships 1-8(Cluster 1) , 8-15(Cluster 2), More than 15(Cluster 0)
2. The maximum of the clusters is contributed by California.

Summary and results of the Algorithms:

Two Class Decision Forest	
True Positive	False Negative
106	14
False Positive	True Negative
29	36
Accuracy	Precision
0.768	0.785
Recall	F1 Score
0.883	0.831
AUC	0.827
Parameters	
Number of decision trees=8	
Maximum depth=16	

Two Class Decision Forest 2	
True Positive	False Negative
113	7
False Positive	True Negative
31	34
Accuracy	Precision
0.795	0.785
Recall	F1 Score
0.942	0.856
AUC	0.856
Parameters	
Number of decision trees=10	
Maximum depth=8	

Two Class Boosted Decision Tree	
True Positive	False Negative
96	24
False Positive	True Negative
26	39
Accuracy	Precision
0.73	0.787
Recall	F1 Score
0.8	0.793
AUC	0.744
Parameters	
Max no of leaves=30	
Max no of samples per leaf node=25	

Two Class Boosted Decision Tree 2	
True Positive	False Negative
101	19
False Positive	True Negative
27	38
Accuracy	Precision
0.751	0.789
Recall	F1 Score
0.842	0.815
AUC	0.769
Parameters	
Max no of leaves=20	
Max no of samples per leaf node=10	

Conclusions and suggestions from clustering:

In our clustering model further analysis will lead to the following insights and help the companies and investors in the following ways

1. The companies which would succeed in each state
2. which factor is prevalent/common among the companies which has a high success rate
3. If a new company is entering the market in a state or an industry what are the preliminary measures it needs to take and brace for the competition.
4. By checking the information of predominant industries and startups in different states, investors and big companies looking for acquisitions can take better business decisions.

Conclusions and suggestions from two class decision forest:

- For a security company, if the number of relationships is less than or equal to 7, the model predicts that the company may not succeed
- In the state of Massachusetts, if the relationships are less than 3 and the average participants are less than one, the company would not succeed.
- If any company with relationships greater than or equal to 4, has venture capitalists and more than one average participant will thrive
- For the state of Missouri, network hosting companies with relationships greater than 3 and average participants more than one has high chances of success.
- For the state of Texas, companies with relationships greater than 3 have been successfully acquired.
- Real estate companies in the state of Colorado and Massachusetts, would not succeed.
- Establishing biotech and mobile companies in North Carolina is not a good idea, due low chances of success
- Advertising companies with more than 7 relationships has a good chance of success.

Note: These are some of the decisions that we would suggest as per our dataset, but these may vary due to other factors in real time.

Pictures of some other insights from the data set:

Sum of relationships	Column Labels																					
Row Labels	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1																						
2										387	410	242	324	325	210	285	288	153	180	152	20	210
3	0	63	158	300	360	490	426	315	328													

Row Labels	22	23	24	25	26	27	28	29	30	31	32	33	35	37	38	42	45	57	63	Grand Total
1	110	115	96	125	78	135	84	58	120	31	32	33	105	111	38	42	45	57	63	1478
2																				3186
3																				2440

Sum of Assignments	Column Labels			
Row Labels	1	2	3	Grand Total
advertising	**5.37%**	**28.19%**	**66.44%**	**100.00%**
CA	9.46%	29.73%	60.81%	100.00%
MA	0.00%	25.00%	75.00%	100.00%
NY	3.45%	34.48%	62.07%	100.00%
other states	0.00%	20.00%	80.00%	100.00%
enterprise	**3.89%**	**27.78%**	**68.33%**	**100.00%**
CA	5.00%	42.50%	52.50%	100.00%
MA	0.00%	30.77%	69.23%	100.00%
NY	20.00%	20.00%	60.00%	100.00%
other states	1.67%	13.33%	85.00%	100.00%
TX	0.00%	11.76%	88.24%	100.00%
mobile	**1.99%**	**27.86%**	**70.15%**	**100.00%**
CA	2.04%	30.61%	67.35%	100.00%
MA	8.70%	26.09%	65.22%	100.00%
NY	0.00%	25.00%	75.00%	100.00%
other states	0.00%	28.00%	72.00%	100.00%
TX	0.00%	0.00%	100.00%	100.00%
other	**1.45%**	**17.97%**	**80.58%**	**100.00%**
CA	1.50%	20.37%	78.13%	100.00%
MA	1.41%	30.99%	67.61%	100.00%
NY	2.22%	22.22%	75.56%	100.00%
other states	0.83%	4.96%	94.21%	100.00%
TX	1.82%	21.82%	76.36%	100.00%
software	**1.21%**	**16.91%**	**81.88%**	**100.00%**
CA	1.52%	18.18%	80.30%	100.00%
MA	0.00%	18.18%	81.82%	100.00%
NY	0.00%	10.00%	90.00%	100.00%
other states	1.85%	14.81%	83.33%	100.00%
TX	0.00%	18.18%	81.82%	100.00%
web	**2.69%**	**21.51%**	**75.81%**	**100.00%**
CA	3.41%	20.49%	76.10%	100.00%
MA	0.00%	14.29%	85.71%	100.00%
NY	1.85%	25.93%	72.22%	100.00%
other states	1.33%	18.67%	80.00%	100.00%
TX	10.00%	60.00%	30.00%	100.00%
Grand Total	**2.07%**	**20.51%**	**77.42%**	**100.00%**

Sum of Assignments	Column Labels			
Row Labels	1	2	3	Grand Total
CA	**2.55%**	**22.81%**	**74.64%**	**100.00%**
advertising	9.46%	29.73%	60.81%	100.00%
enterprise	5.00%	42.50%	52.50%	100.00%
mobile	2.04%	30.61%	67.35%	100.00%
other	1.50%	20.37%	78.13%	100.00%
software	1.52%	18.18%	80.30%	100.00%
web	3.41%	20.49%	76.10%	100.00%
MA	**1.38%**	**23.96%**	**74.65%**	**100.00%**
advertising	0.00%	25.00%	75.00%	100.00%
enterprise	0.00%	30.77%	69.23%	100.00%
mobile	8.70%	26.09%	65.22%	100.00%
other	1.41%	30.99%	67.61%	100.00%
software	0.00%	18.18%	81.82%	100.00%
web	0.00%	14.29%	85.71%	100.00%
NY	**2.57%**	**23.53%**	**73.90%**	**100.00%**
advertising	3.45%	34.48%	62.07%	100.00%
enterprise	20.00%	20.00%	60.00%	100.00%
mobile	0.00%	25.00%	75.00%	100.00%
other	2.22%	22.22%	75.56%	100.00%
software	0.00%	10.00%	90.00%	100.00%
web	1.85%	25.93%	72.22%	100.00%
other states	**1.06%**	**12.39%**	**86.55%**	**100.00%**
advertising	0.00%	20.00%	80.00%	100.00%
enterprise	1.67%	13.33%	85.00%	100.00%
mobile	0.00%	28.00%	72.00%	100.00%
other	0.83%	4.96%	94.21%	100.00%
software	1.85%	14.81%	83.33%	100.00%
web	1.33%	18.67%	80.00%	100.00%
TX	**1.82%**	**21.82%**	**76.36%**	**100.00%**
enterprise	0.00%	11.76%	88.24%	100.00%
mobile	0.00%	0.00%	100.00%	100.00%
other	1.82%	21.82%	76.36%	100.00%
software	0.00%	18.18%	81.82%	100.00%
web	10.00%	60.00%	30.00%	100.00%
Grand Total	**2.07%**	**20.51%**	**77.42%**	**100.00%**

Count of category_code	Column Labels						
Row Labels	advertising	enterprise	mobile	other	software	web	Grand Total
CA	6.78%	7.19%	8.01%	46.41%	15.20%	16.43%	100.00%
MA	7.23%	6.02%	12.05%	33.73%	28.92%	12.05%	100.00%
NY	11.32%	4.72%	8.49%	49.06%	6.60%	19.81%	100.00%
other states	5.39%	10.78%	9.31%	41.18%	19.61%	13.73%	100.00%
TX	0.00%	14.29%	4.76%	50.00%	19.05%	11.90%	100.00%
Grand Total	**6.72%**	**7.92%**	**8.57%**	**44.58%**	**16.59%**	**15.62%**	**100.00%**

THANK YOU