INFINITY STUDIO
shape your imagination

Deshpande Startups

Investment Predictor Model

Submitted by

Ayush Gupta

Trinity College Dublin

Ayushgupta.w19@gmail.com

ABSTRACT

This project makes use of the dataset about start-ups in Indian Ecosystem framed as a classification problem, predicting whether start-ups can raise a certain threshold to predict performance. In this project the threshold has been decided as the ability of the company to raise Series-A Funding (1.25M to 7M USD). Comparisons of P-value, Chi-square, F-value, and Mutual Information as variable selection methods are presented and evaluated. Finally, the chosen model finds the most important predictors: years established, Location such as: Bangalore, Delhi, Mumbai, NCR and Industry Verticals: eCommerce, health-tech, fintech, ed-tech etc. Most of these are intuitive as they all confirm the geographical advantages and market competitiveness effects on start-ups.

Problem Statement

Predicting the probability of a company in the Indian start-up ecosystem which has received seed funding and pre-Series-A funding to reach Series-A funding. In this problem we have defined ranges for Series A, pre-Series A and Seed Funding which are as follows:

- Seed Funding: up-to 500k USD
- Pre-Series A: 500k USD to 1.5M USD
- Series A: 1.5M to 7M USD

Solution

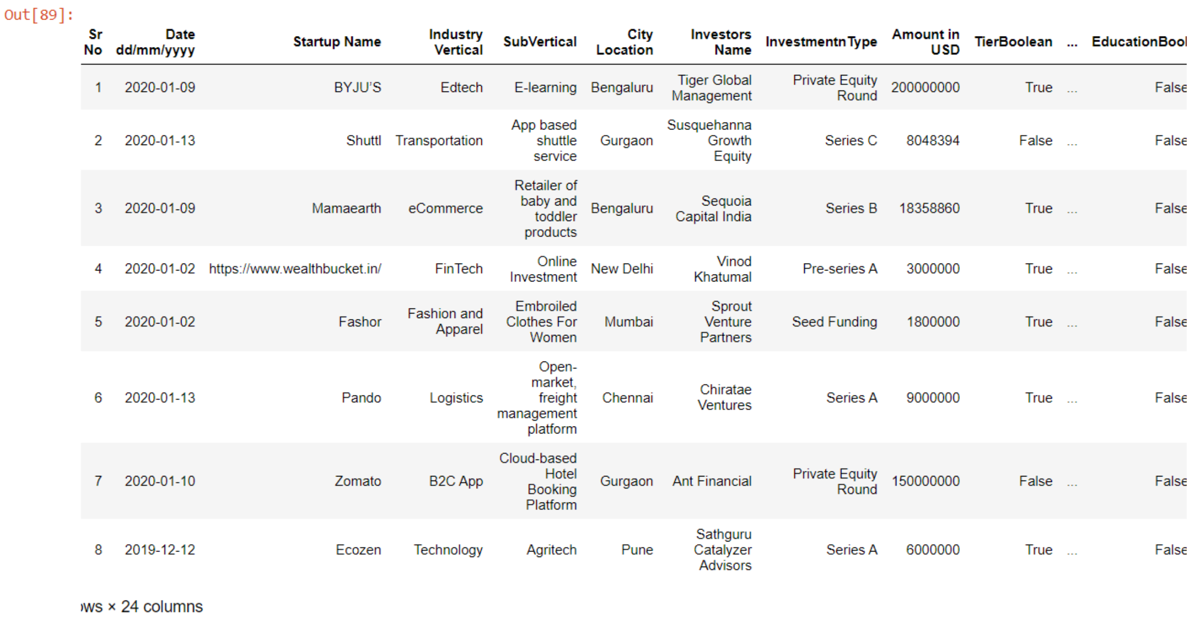
➢ Data Acquisition:

After researching a few different start-up databases, I chose the database from Kaggle for its detailed data. Along with that I collected data from angelist.co by writing a script which would allow me to scrape data. I used both the datasets to create a master sheet which had all the data of the startups in the Indian ecosystem.

The following data is presented in the dataset:

- Date of Receiving funding
- Location of startup
- The Industry Vertical
- Current stage of the startup
- Kind of investment raised (seed/ angel/ pre-seriesA, etc)
- Amount raised in USD

This is how the final dataset looks like:



Out[89]:

Sr No	Date dd/mm/yyyy	Startup Name	Industry Vertical	SubVertical	City Location	Investors Name	InvestmentnType	Amount in USD	TierBoolean	...	EducationBool
1	2020-01-09	BYJU'S	Edtech	E-learning	Bengaluru	Tiger Global Management	Private Equity Round	200000000	True	...	False
2	2020-01-13	Shuttl	Transportation	App based shuttle service	Gurgaon	Susquehanna Growth Equity	Series C	8048394	False	...	False
3	2020-01-09	Mamaearth	eCommerce	Retailer of baby and toddler products	Bengaluru	Sequoia Capital India	Series B	18358860	True	...	False
4	2020-01-02	https://www.wealthbucket.in/	FinTech	Online Investment	New Delhi	Vinod Khatumal	Pre-series A	3000000	True	...	False
5	2020-01-02	Fashor	Fashion and Apparel	Embroiled Clothes For Women	Mumbai	Sprout Venture Partners	Seed Funding	1800000	True	...	False
6	2020-01-13	Pando	Logistics	Open-market, freight management platform	Chennai	Chiratae Ventures	Series A	9000000	True	...	False
7	2020-01-10	Zomato	B2C App	Cloud-based Hotel Booking Platform	Gurgaon	Ant Financial	Private Equity Round	150000000	False	...	False
8	2019-12-12	Ecozen	Technology	Agritech	Pune	Sathguru Catalyzer Advisors	Series A	6000000	True	...	False

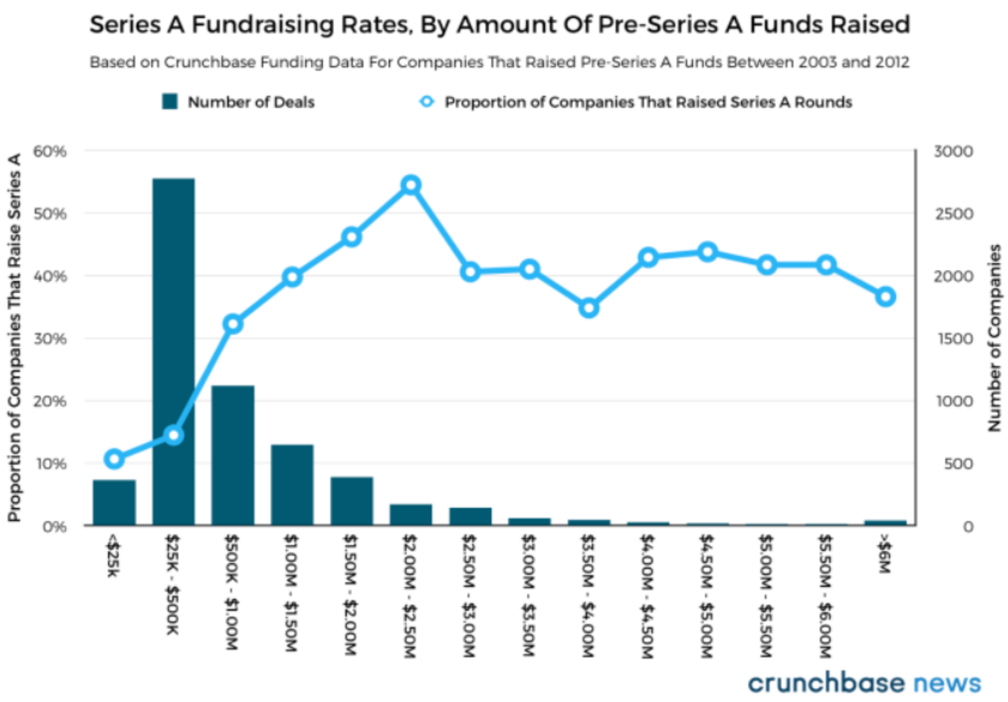
ws × 24 columns

➢ Evaluation Metric

With the data that we have, it is very difficult to decide if we wish to use a classification between “not good” and “good” startups. If we base this problem as a classification problem, we would need a good measure of what is good. I was inspired by a study by TechCrunch regarding the fundraising impact of one round on the next.

We see in the chart below, the probability of raising Series A is highest when companies raised $2 million to $2.5 Million in pre-series A. Therefore, a seed-stage company that raised $2M should have a higher chance of survival than one that raised less than $2M.

As the data is scarce about the progression of companies reaching various stages of funding, I decided to use reaching_seriesA as a threshold to evaluate the companies.



Series A Fundraising Rates, By Amount Of Pre-Series A Funds Raised

Based on Crunchbase Funding Data For Companies That Raised Pre-Series A Funds Between 2003 and 2012

crunchbase news

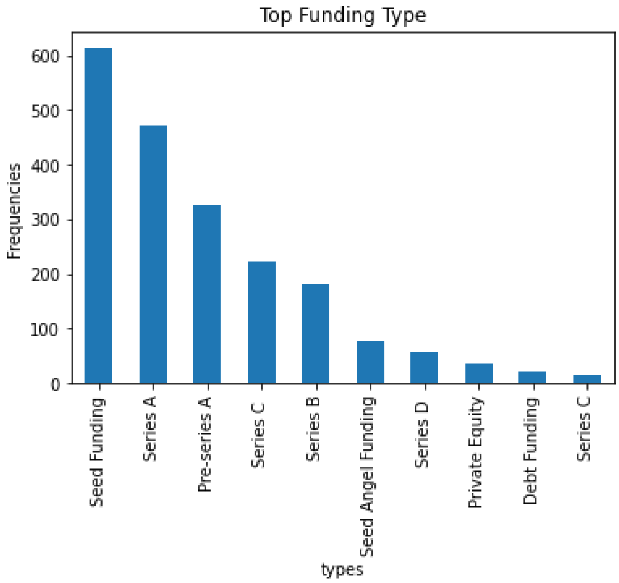
➢ Data Pre-processing

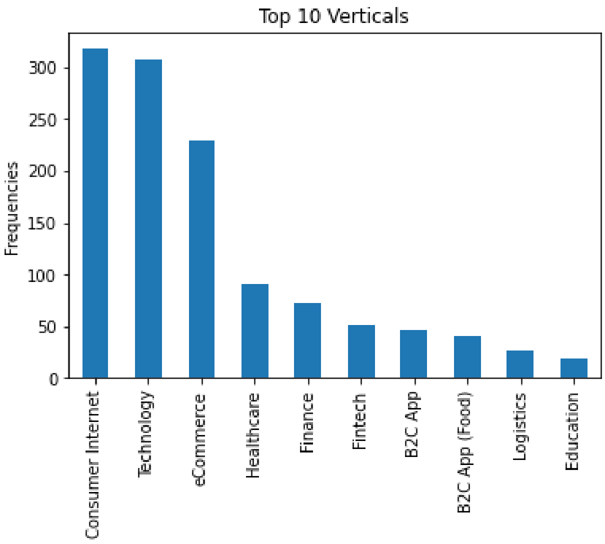
After correcting the data types and cleaning repeated rows/ special characters, the following processing logic is applied to create a table ready for training.

- The raw data columns “City Location” and “Industry Vertical” are one hot encoded for easier evaluation in machine learning model.
- Filtered stage by “seed” and “Series A” and had two datasets ready for modelling.
- The raw data columns of “InvestmentnType” was modified to change the private equity classification into seed funding, pre-SeriesA, Series A funding according to this:
  - Seed: upto $500K
  - Pre-SeriesA: $500K to $1.5M
  - Series A: $1.5M to $7M
- Changed “Industry Vertical” column into less ambiguous classification buckets

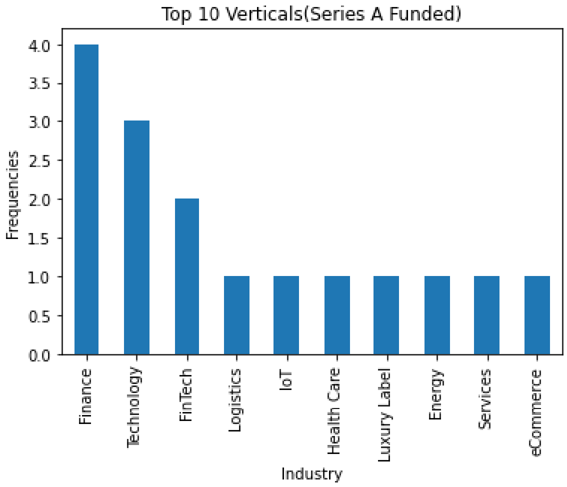
➢ Preliminary Data Visualization

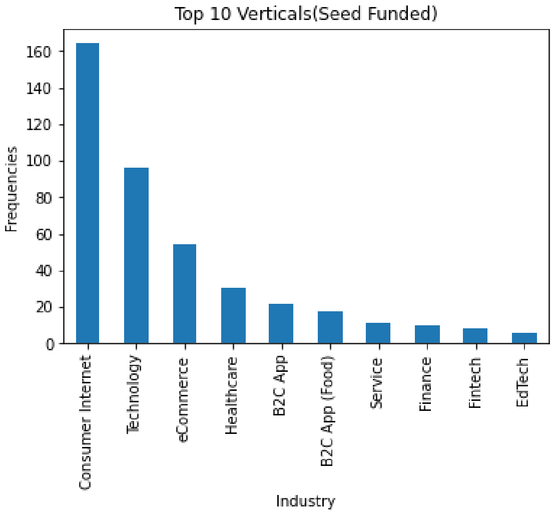
Our dataset contained a total of 3042 companies from all different stages, here are two interesting charts that show the most-raised(famous) startups as well as the distribution of stages.



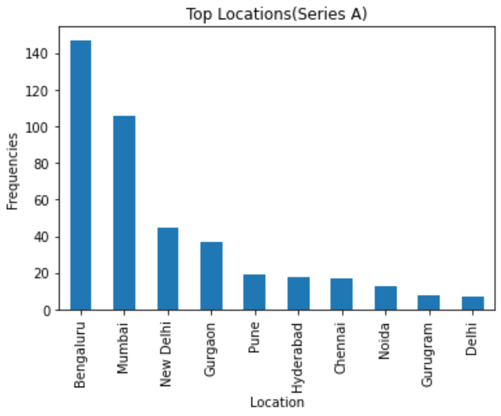


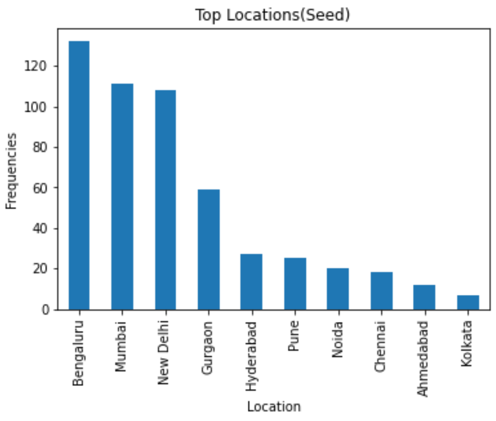
Now, going into specifically seed and series A. let's see how seed-stage companies and series A companies compare in terms of their distribution in location and industry verticals

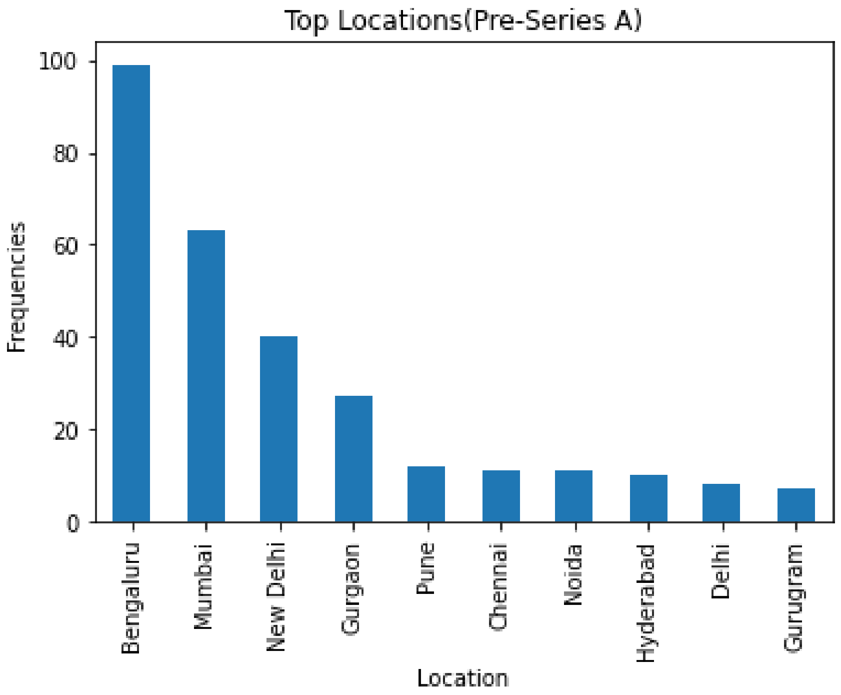




The hottest markets seem to be somewhat consistent between seed and Series A. Finance and Technology are the top of the leader board.

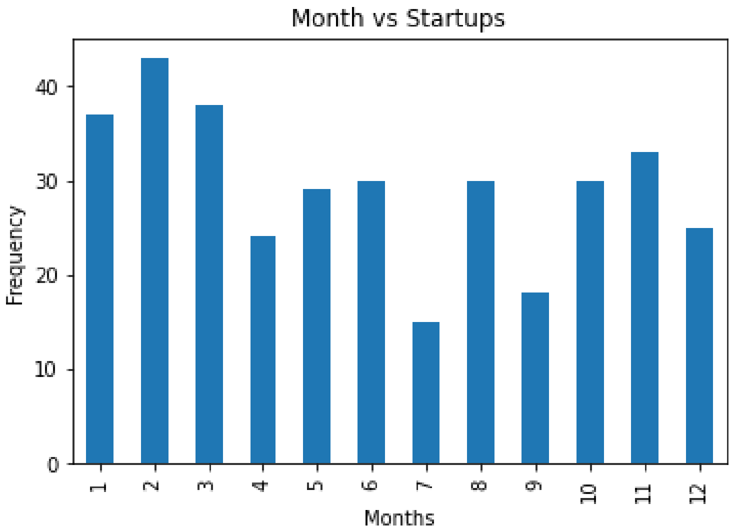


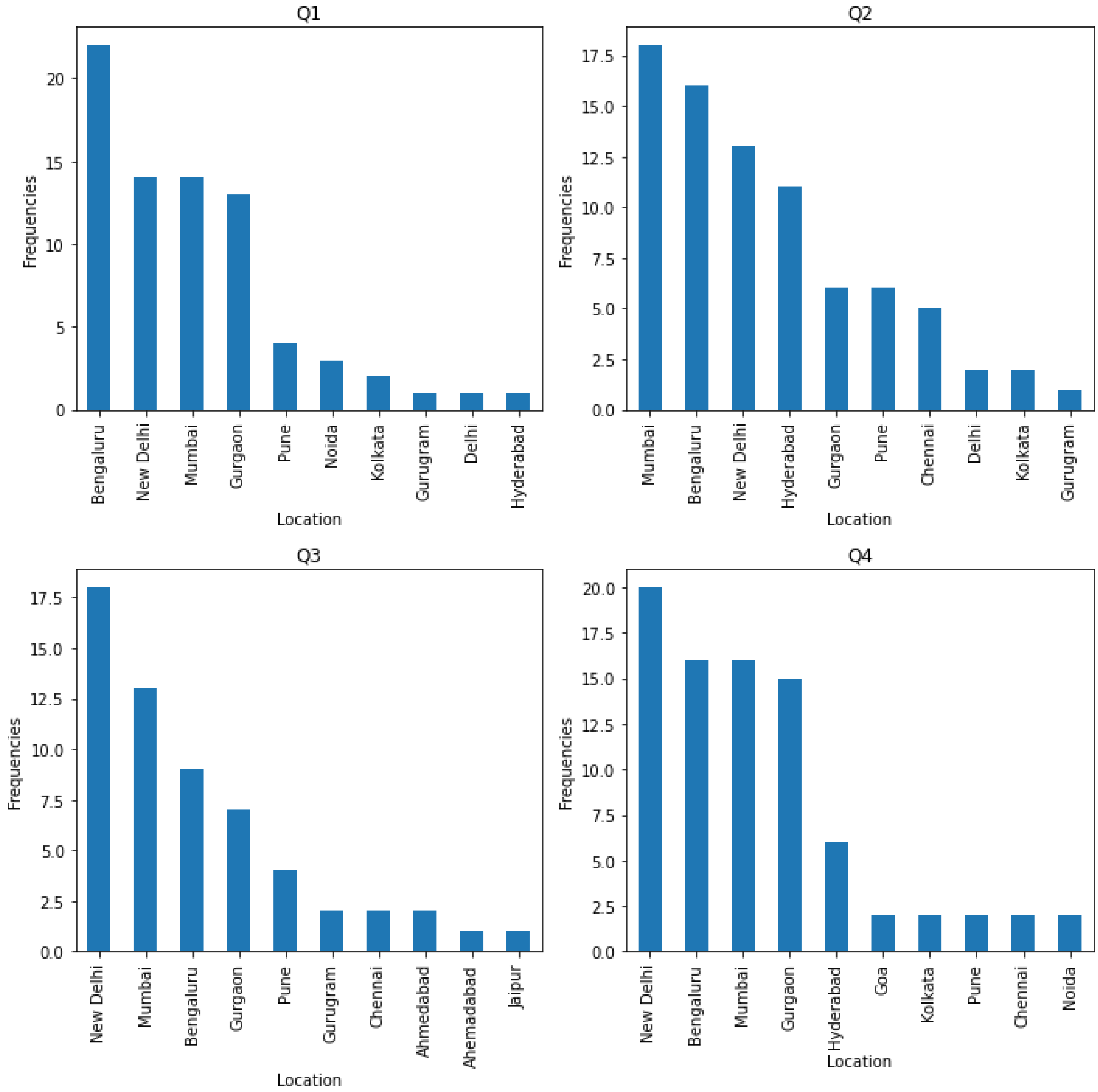




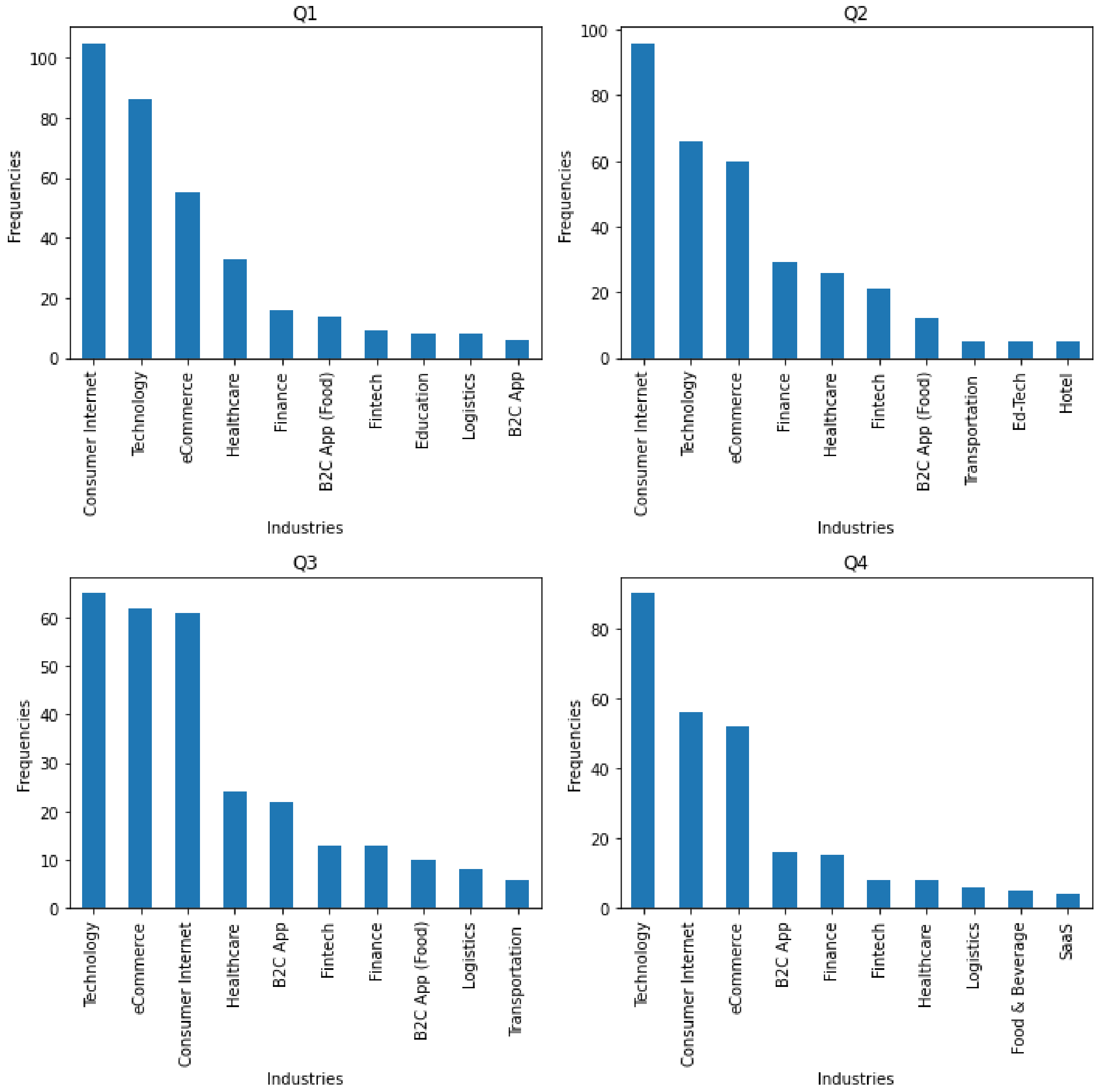
Locations also overlap between seed and series A. We can see Bengaluru, Mumbai and New Delhi are popular homes for all the 3 stages.

I have included some other interesting graphs which can deepen our understanding about how the different parameters interact with each other.

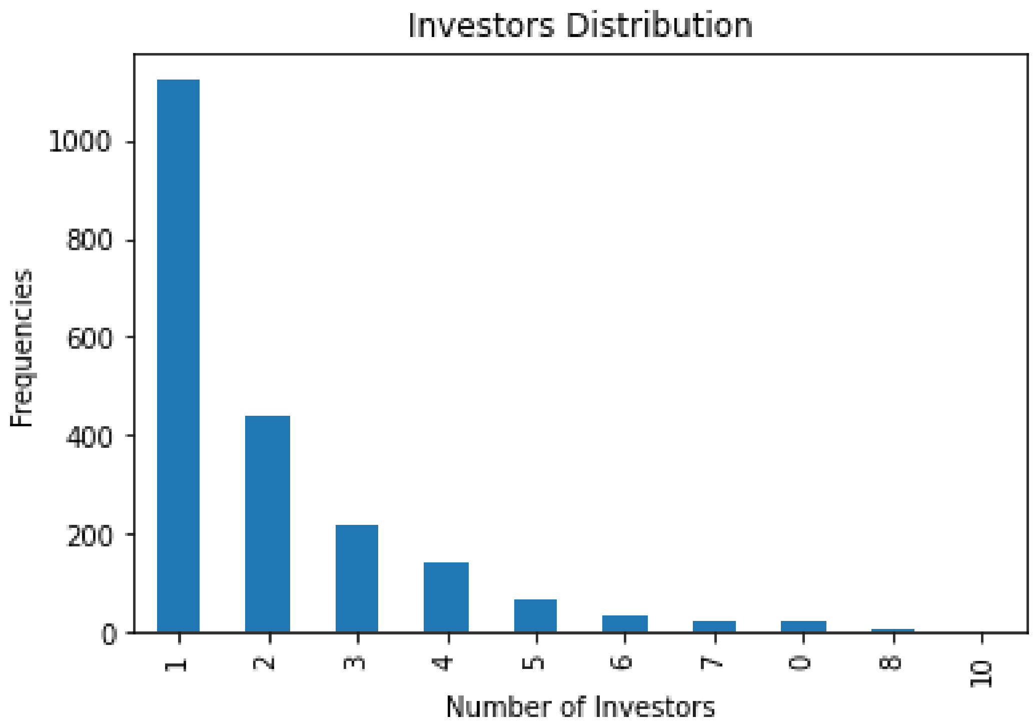




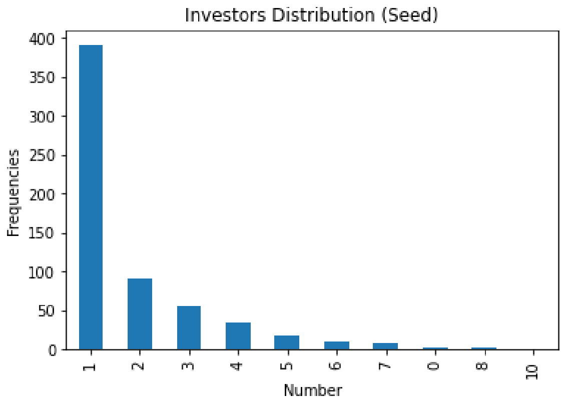
The quarter wise distribution of location in which the startups pop up are shown above. Its interesting to see how New Delhi becomes a hot market for startups in Q3 and Q4. (what do you think the reason can be?)

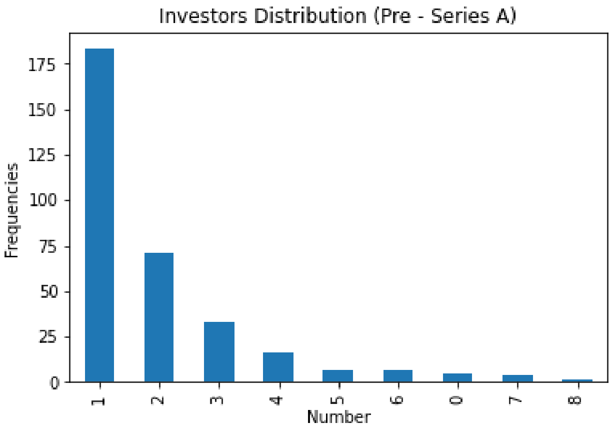


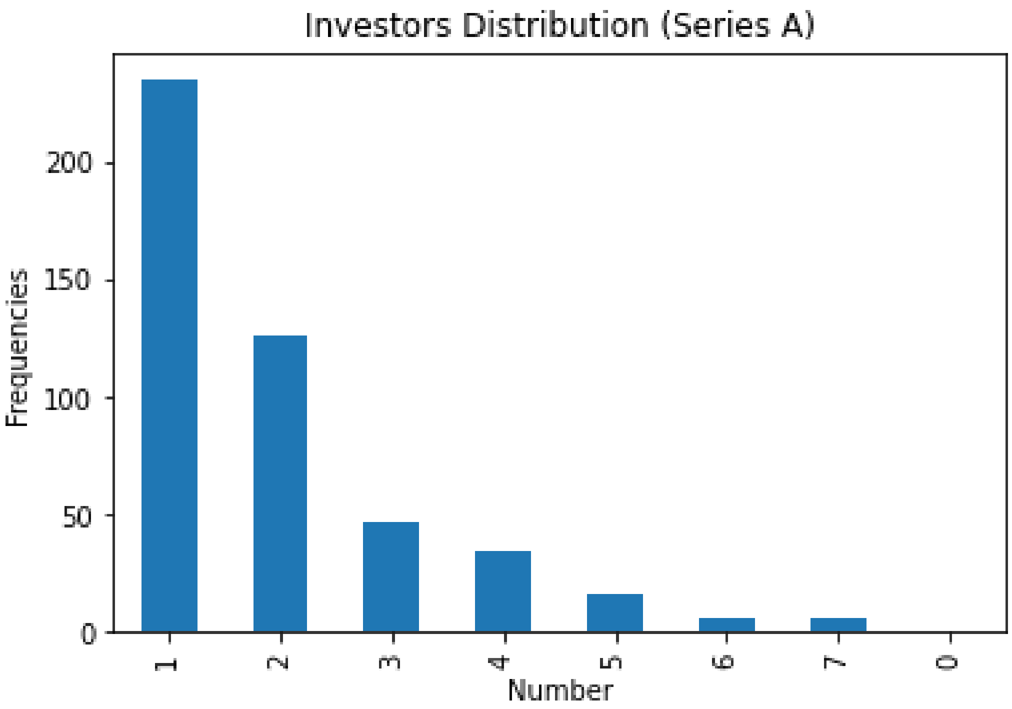
Similarly we can see the quarter-wise distribution of industries. We see that Technology and Consumer Internet are consistently the industries which create the most startups, we see very little fluctuation in this space.



We can see the number of investors which invest in a company. Startups generally have only 1 investor and while some might have more we can see in the next graph how even at different stages of startups, they're mostly supported by 1 investor.







➢ Modelling

If we inspect the nature of the input features, we find most of them are binary ones that are generated by one-hot encoding. Our input data has high dimensionality (and most are binary), hence not very suitable for tree-based models. Given the sparse distribution and low complexity of the data, I chose to use **Logistic Regression** to model it. (KNNs can give good results as well).

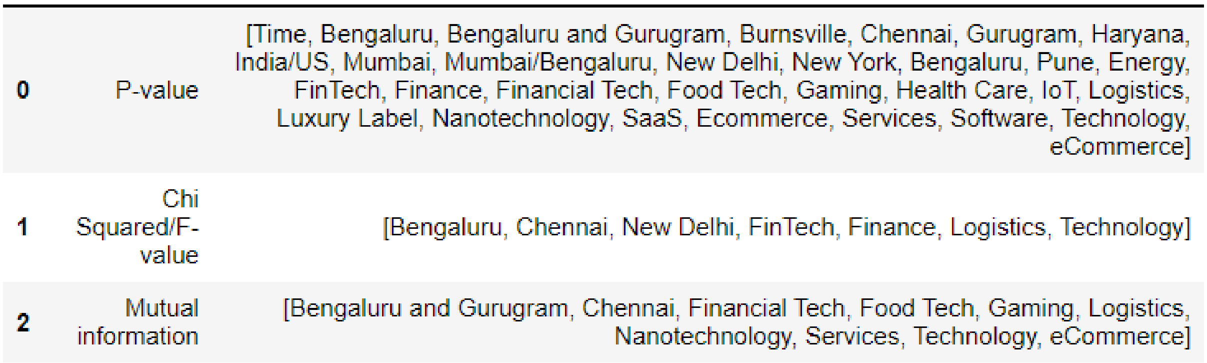
➢ Feature Selection

The binary features can be very skewed towards 0. (E.g. one company has a location of “Haridwar”, then a column called “Haridwar” where exactly one row has a value of 1 where the rest are all 0s.) So certain variable would be comparatively less important than others. In this step, feature selection will be used to select only the features that are important.

I had decided to try a couple of different feature selection methods and see how they influence the result. Namely, I will use *Chi-Squared, mutual information, ANOVA f-value, and p-value*

For the p value selection, I performed a backward elimination using an alpha of 0.05 and the rest are utilizing built in functions from ***sklearn.feature_selection***.

The p-value backward elimination returns a non-fixed number of variables in this case, 29 for Seed. The sci-kit learn selections are implemented using SelectKBest, where K is set to 10. This number is picked by finding the best validation set AUC performance. Below are the choices of variables selected using the different processes.

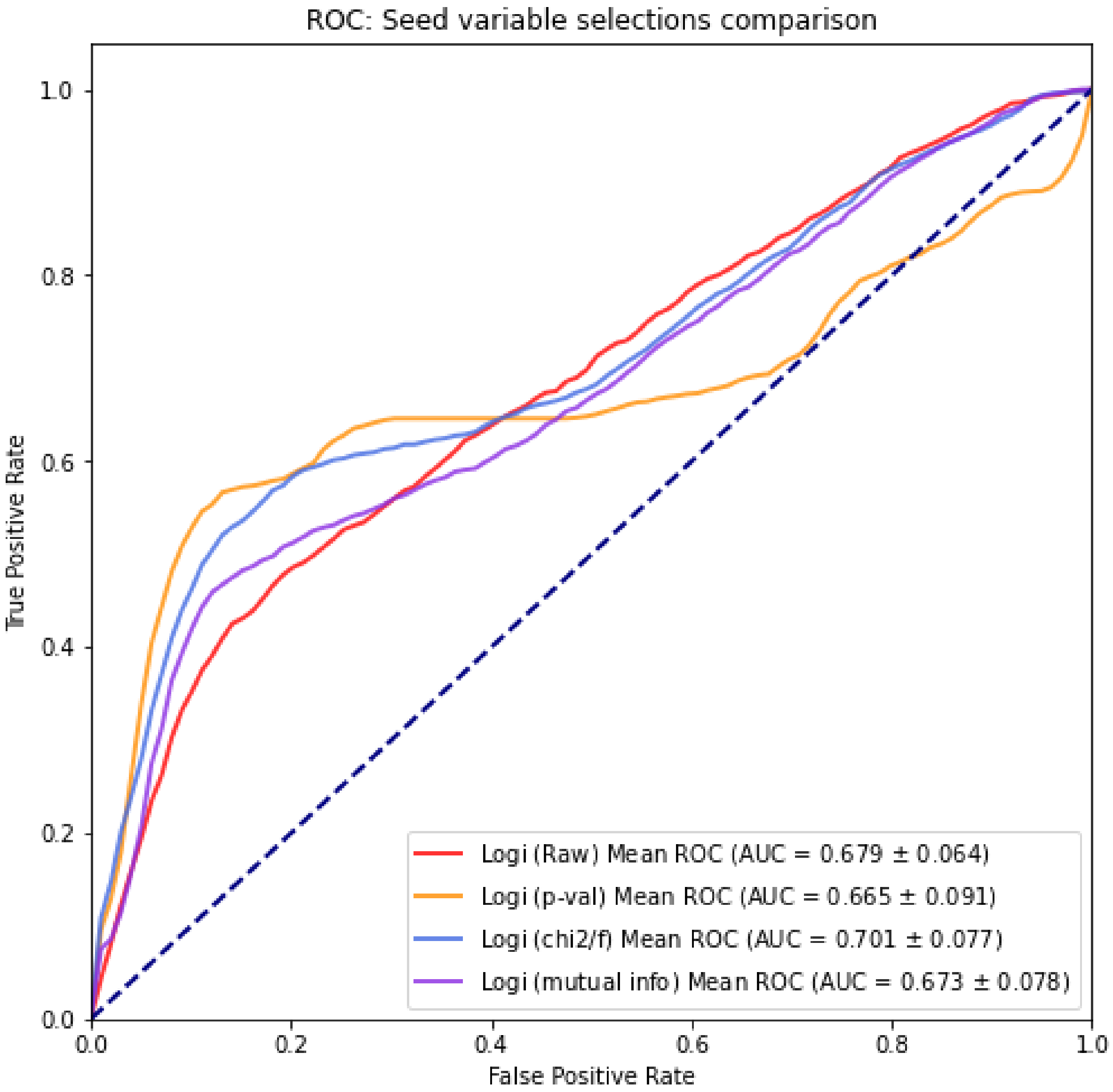


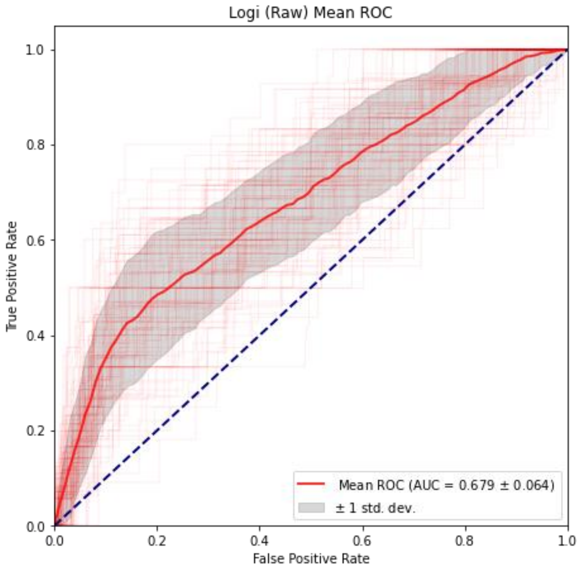
0	P-value	[Time, Bengaluru, Bengaluru and Gurugram, Burnsville, Chennai, Gurugram, Haryana, India/US, Mumbai, Mumbai/Bengaluru, New Delhi, New York, Bengaluru, Pune, Energy, FinTech, Finance, Financial Tech, Food Tech, Gaming, Health Care, IoT, Logistics, Luxury Label, Nanotechnology, SaaS, Ecommerce, Services, Software, Technology, eCommerce]
1	Chi Squared/F-value	[Bengaluru, Chennai, New Delhi, FinTech, Finance, Logistics, Technology]
2	Mutual information	[Bengaluru and Gurugram, Chennai, Financial Tech, Food Tech, Gaming, Logistics, Nanotechnology, Services, Technology, eCommerce]

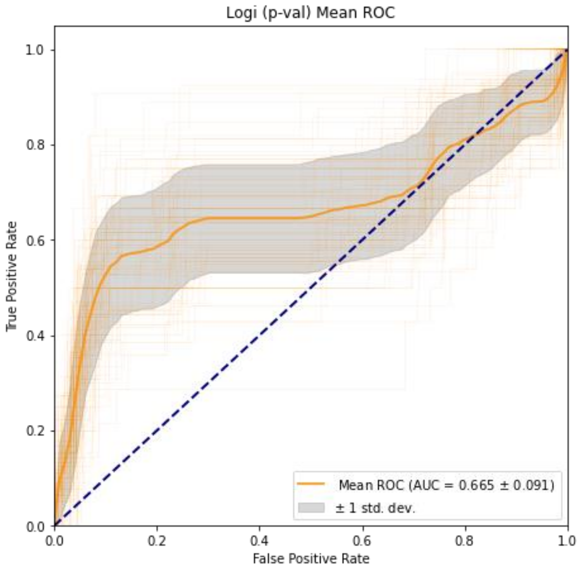
➢ Model Performances

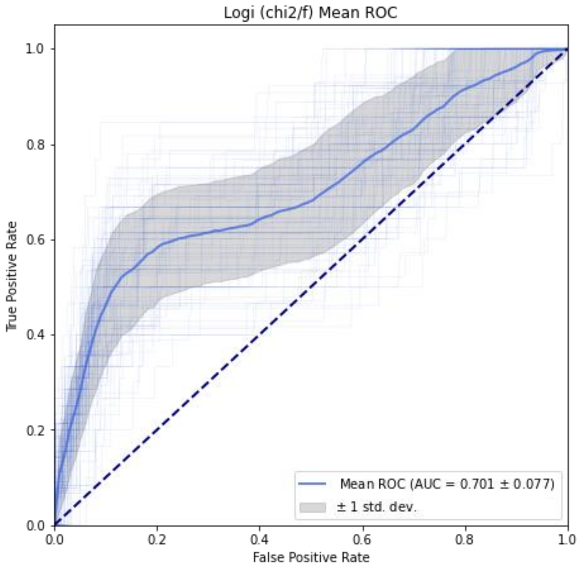
For each of the feature selection methods, logistic regression is trained on that set of features, and a ROC curve is plotted using validation sets. Each ROC curve has metric AUC (Area Under the Curve). The more this value leans to 1, the better the model at predicting the outcome correctly.

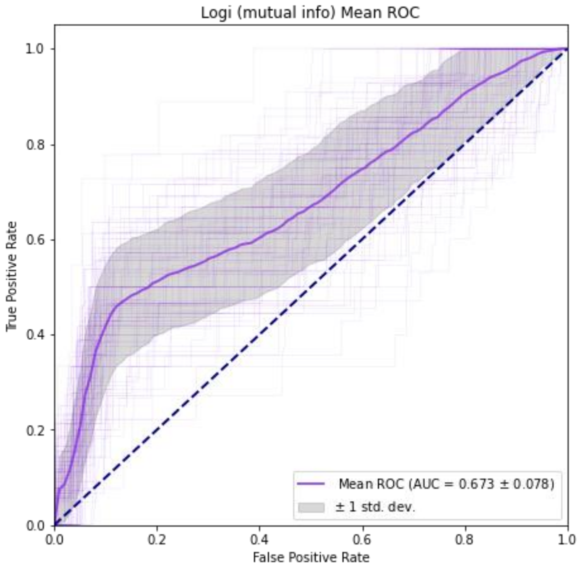
Cross-validation is used in this step to yield a smooth and rounded "Mean" ROC curve. Each faint curve in the 4x-grid is a single ROC from a train/validation set, the bolded curve represents the mean ROC, with a shaded region in grey representing one standard deviation. The large plot is simply the mean ROC plotted together.







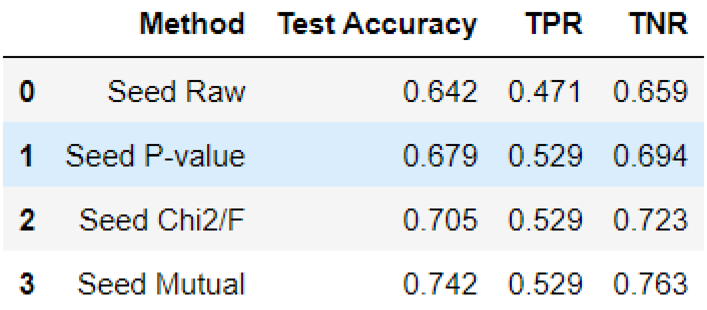




For Seed companies, the **Chi-squared/F-Value** selection method performs the best on the validation set with an AUC of **0.701**

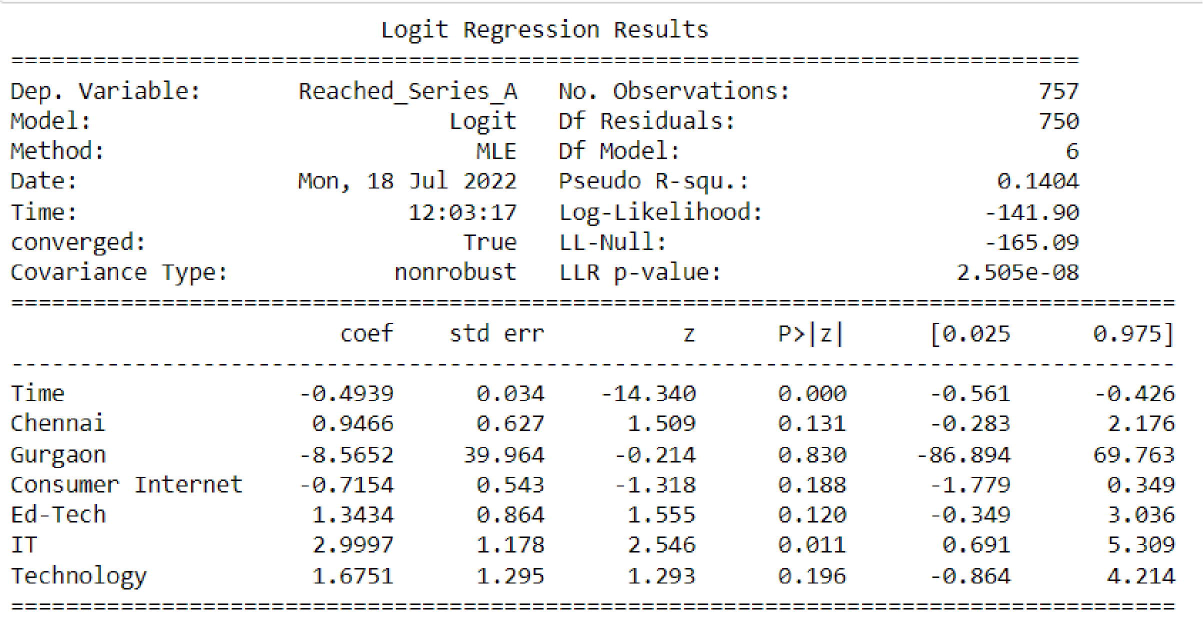
➢ Model Evaluation and Discussion

For Seed stages, indeed the **Mutual information** selection gave us the best test accuracy of **0.742.** This means we can correctly predict the ability to raise Series A **0.742** of the time. The TPR and TNR also are reasonable. Although raising Series A does not infer a definite success, it has the most probability of raising the next round. So, it can definitely help us identify “strong” start-ups.



	Method	Test Accuracy	TPR	TNR
0	Seed Raw	0.642	0.471	0.659
1	Seed P-value	0.679	0.529	0.694
2	Seed Chi2/F	0.705	0.529	0.723
3	Seed Mutual	0.742	0.529	0.763

Let us visit again the variables identified by the Chi2/F selection and their corresponding coefficients and p-values. This will provide us insight on what factors are strong influencers of the survival of a startup.



```
                           Logit Regression Results                           
==============================================================================
Dep. Variable:        Reached_Series_A   No. Observations:                  757
Model:                           Logit   Df Residuals:                      750
Method:                            MLE   Df Model:                            6
Date:                 Mon, 18 Jul 2022   Pseudo R-squ.:                  0.1404
Time:                         12:03:17   Log-Likelihood:                -141.90
converged:                        True   LL-Null:                       -165.09
Covariance Type:             nonrobust   LLR p-value:                 2.505e-08
=====================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Time               -0.4939      0.034    -14.340      0.000      -0.561      -0.426
Chennai             0.9466      0.627      1.509      0.131      -0.283       2.176
Gurgaon            -8.5652     39.964     -0.214      0.830     -86.894      69.763
Consumer Internet  -0.7154      0.543     -1.318      0.188      -1.779       0.349
Ed-Tech             1.3434      0.864      1.555      0.120      -0.349       3.036
IT                  2.9997      1.178      2.546      0.011       0.691       5.309
Technology          1.6751      1.295      1.293      0.196      -0.864       4.214
=====================================================================================
```

Outcomes

Overall, the quality of data is certainly limited, but our seed-stage model performed well with an accuracy of **0.742.** The predictors of a successful start-up are far more than just the Industry Vertical, location, and years since they have received capital. Factors like company culture and founder's background may be more deterministic in a real-life situation.

Nevertheless, we got significant insights from the seed stage model that is very applicable to real-life knowledge. So as a takeaway, put your money in a Delhi, Gurugaon, fintech and Technology.

Taking a step back, this project looks at a very simplified problem based on our understanding of the Bharat ecosystem and information about the startups in the public domain. There is so much more mystery, science, and techniques in the VC realm that simply might not be learnable by logistic regression.

References:

1. Building a Logistic Regression model: https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

2. Dataset aggregated from Startup India: https://www.startupindia.gov.in/

3. Indian startup ecosystem Dataset:

https://www.kaggle.com/datasets/sudalairajkumar/indian-startup-funding

4. Patterns of a successful startup: A study by Techcrunch.

https://techcrunch.com/2017/08/23/does-it-really-matter-how-much-your-startup-raises/