

A Project Report  
on  
**PREDICTIVE APPLICATION ON  
START-UPS ACQUISITION STATUS PREDICTION**

*Work carried out for*  
**Technocolabs Softwares Pvt. Ltd., Indore, M.P.**



*Submitted to Pondicherry University in partial fulfilment of the requirement for  
the Award of the Degree of*

***Master of Business Administration***

By

**MARIA JAMES**

(Reg. No. 21401023)

*Under the Supervision of*

**Dr. L. MOTHILAL, Associate Professor**

Department of Management Studies, Pondicherry University

&

**YASIN SHAH, CEO,**

Technocolabs Softwares Pvt. Ltd.



**Department of Management Studies**

Pondicherry University, Pondicherry, INDIA – 605 014

July – September 2022

## DECLARATION

I hereby declare that the project titled, “**Predictive Application on Start-ups Acquisition Status Prediction**” is original work done by me under the guidance of **Dr. L. Mothilal, Associate Professor**, Department of Management Studies, Pondicherry University, and **Yasin Shah, CEO, Technocolabs Pvt. Ltd.** This project or any part thereof has not been submitted for any Degree / Diploma / Associateship / Fellowship / any other similar title or recognition to this University or any other University.

I take full responsibility for the originality of this report. I am aware that I may have to forfeit the degree if plagiarism has been detected after the award of the degree. Notwithstanding the supervision provided to me by the Faculty Guide, I warrant that any alleged act(s) of plagiarism in this project report are entirely my responsibility. Pondicherry University and/or its employees shall under no circumstances whatsoever be under any liability of any kind in respect of the aforesaid act(s) of plagiarism.



**Maria James**

21401023

II MBA Data Analytics

Department of Management Studies

Pondicherry University

Place: Pondicherry

Date: 28 October 2022

## **Acknowledgements**

*First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my internship to complete the project successfully.*

*I would like to express my sincere gratitude to my research guide Dr. L. Mothilal, Associate Professor, Department of Management Studies, Pondicherry University for his incessant encouragement and support extended throughout the research period. He guided me throughout the project and taught me how to present the project as clearly as possible.*

*Thanks to Mr. Yasin Shah, Technocolabs Softwares, for sharing and discussion of various research related matters.*

*Many thanks to those faculty members who helped me in sharpening my thinking by cheerfully providing challenging comments and questions.*

*Non-teaching staff of the Department of Management Studies, Pondicherry University was extremely helpful to me during the research period and therefore, I am thankful to them.*

*I thank my classmates of Pondicherry University who had given me moral support.*

*I am extremely grateful to my family who had to tolerate late night work, and curtailed weekends & vacations. The support and encouragement of my parents gave me the energy, stamina, and inspiration to complete the project.*

**Maria James**



## DEPARTMENT OF MANAGEMENT STUDIES

Pondicherry University, Pondicherry – 605 014, India

Ph: (O) 91-413-2654503, Mobile:99947-38415

---

**Dr. L. MOTHILAL**, M.B.A., Ph.D.  
*Associate Professor*

### C E R T I F I C A T E

This is to certify that the project titled “**Predictive Application on Start-ups Acquisition Status Prediction**” submitted for the award of Degree of Master of Business Administration, in the Department of Management Studies, Pondicherry University, Pondicherry, India, is a record of bonafide project work carried out by Ms. Maria James under my supervision.

  
**(L. MOTHILAL)**

**Dr. B. Charumathi**

*Professor & Head,*

Department of Management Studies,  
Pondicherry University



**Dr. L. MOTHILAL**, M.B.A., Ph.D.  
*Associate Professor*

Department of Management Studies  
Pondicherry University  
Pondicherry - 605 014, India  
Ph(O) 0413-2654503, Mob: 99947-38415  
Email: [mothilal@pondiuni.ac.in](mailto:mothilal@pondiuni.ac.in)

Place: Pondicherry

Date: 28 October 2022

# PROJECT COMPLETION LETTER



Date : 20-09-2022

Dear Sir/Madam,

This is to certify that Ms. Maria James has completed an internship program from **15th July 2022 to 15th September 2022** at Technocolabs, Indore. During this internship, we found her to be punctual, hardworking, and inquisitive. She worked on a Data Analysis project for the company on various domains of tasks such as Data Analysis, Data Manipulations, Data Classification techniques, Data Visualization, and Data Modelling with Deployment on Cloud Services. She developed the project and completed it within the given deadline.

She has worked on various tasks on the final project on the **Predictive Application on Startup Acquisition Status Prediction** under the mentorship and guidance of **Mr. Yasin Shah**.

Best wishes,

A handwritten signature in dark blue ink that reads 'Yasin'.

Yasin Shah  
Founder & CEO Technocolabs

# CONTENTS

Sl. No.	Particulars	Page No.
	<i>Declaration</i>	
	<i>Acknowledgments</i>	
	<i>Certificate from University</i>	
	<i>Certificate from Organization</i>	
	<i>Acronyms</i>	
	<i>Figures</i>	
<b>Chapter - I</b>	<b>Introduction – Start-up: An Overview</b>	10
1.1	Introduction	11
1.2	The scenario of Start-ups Status	11
1.3	Need for the Study	12
1.4	Objectives of the Study	12
<b>Chapter - II</b>	<b>Research Methodology</b>	13
2.1	Source of Data	14
2.2	Summary of the Data	14
2.3	Statistical Tools Used	15
2.4	Softwares Used	16
2.5	Scope of the Study	17
2.6	Significance of the Study	17
2.7	Limitations	17

# CONTENTS

Sl. No.	Particulars	Page No.
<b><i>Chapter-III</i></b>	<b>Company Profile and Industry Profile</b>	18
3.1	Company Profile	19
3.2	Industry Profile	20
<b><i>Chapter-IV</i></b>	<b>Data Analysis and Findings</b>	22
4.1	Data Pre-Processing	23
4.2	Exploratory Data Analysis	25
4.3	Feature Engineering	33
<b><i>Chapter-V</i></b>	<b>Model Building</b>	37
5.1	Model Building	38
5.2	Model Validation	40
5.3	Predictor Function	42
<b><i>Chapter-VI</i></b>	<b>Model Deployment</b>	43
6.1	Model Deployment	44
<b><i>Chapter-VII</i></b>	<b>Suggestions &amp; Conclusion</b>	46
7.1	Suggestions	47
7.2	Conclusion	47
<b><i>References</i></b>		

## ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
DL	Deep Learning
EC2	Elastic Cloud Compute
EDA	Exploratory Data Analysis
IPO	Initial Public Offering
JS	Java Script
MI	Mutual Information
ML	Machine Learning
QDA	Quadratic Discriminant Analysis
RF	Random Forest
S3	Simple Storage Service
SaaS	Software as a Service
PaaS	Platform as a Service
IDE	Integrated development environment
GUI	Graphical User Interface
NumPy	Numerical Python



## Figures

<b>Sl. No.</b>	<b>Particulars</b>	<b>Page No.</b>
4.2.1	<i>Distribution of Companies based on its Functional Status</i>	25
4.2.2a	<i>Top Categories</i>	36
4.2.2b	<i>Statuses of Top 5 Categories</i>	26
4.2.3	<i>Company Categories with less than 5 years life</i>	27
4.2.4a	<i>Categories – Operating</i>	27
4.2.4b	<i>Origination Year – Operating</i>	28
4.2.5a	<i>Categories – IPO</i>	28
4.2.5b	<i>Origination Year – IPO</i>	29
4.2.6a	<i>Categories – Acquired</i>	29
4.2.6b	<i>Origination Year – Acquired</i>	30
4.2.6c	<i>Closing Year – Acquired</i>	30
4.2.7a	<i>Categories – Closed</i>	31
4.2.7b	<i>Origination Year – Closed</i>	31
4.2.7c	<i>Closing Year – Closed</i>	32
4.2.3	<i>Start-ups Location vs Status</i>	32
4.3.1	<i>Mutual Information Scores</i>	34
4.3.2	<i>ExtraTreesClassifier Scores</i>	35
4.3.3	<i>Standardization</i>	36
5.1.1	<i>QDA – Prediction</i>	39
5.1.2	<i>Random Forest Prediction</i>	40
5.2.1	<i>Confusion Matrix for QDA</i>	41
5.2.2	<i>Confusion Matrix for RF</i>	41
5.2.3	<i>Classification Report for QDA</i>	41
5.2.4	<i>Classification Report for RF</i>	42
5.3.1	<i>Predictor Function Prediction</i>	42

## *Chapter – I*

---

# *Introduction – Start-up: An Overview*

---

## **1.1 Introduction**

Start-ups are companies or organizations developed by one or more entrepreneurs with an objective of developing a particular product or service that is in demand. It can be small business start-ups that are mainly self-funded and that grow at their own pace or big business start-ups who come into picture with big investments and run along with the competitors from the beginning.

Google, Facebook, Twitter, etc were start-ups that has become inevitable in the modern era.

## **1.2 Scenario of Start-ups Status**

Start-ups that have a good product-market fit, with small test markets and passionate about disruption are likely to survive long duration in the market.

A company pursue an IPO for raising the capital, to obtain publicity, etc. In an IPO, a privately owned company lists its shares on a stock exchange, making them available for purchase by the general public.

When a new start-up in the market competes with the existing big-shots, the later buys all or part of another company's stock or assets to gain control of and expand on the target company's strengths.

But, being in the wrong market, a lack of research, bad partnerships, ineffective marketing and not being an expert in the industry can lead to the failure of any venture.

It is at this point the importance of this project comes into picture.

### **1.3 Need for the Study**

Through this project, it is aimed to predict the status of a start-up, if it is in ‘operating’, ‘acquired’, ‘IPO’ or ‘closed’ statuses based on its financial information.

Year of foundation, 1<sup>st</sup> and last funding year, funding rounds, total funding, 1<sup>st</sup> and last milestones and total milestones achieved are the financial variables used for predicting the company’s status.

The data is cleaned and feature engineering techniques are applied before modelling and predicting. After successful model validation, the model is used to create an app in Streamlit and is then deployed to Heroku.

### **1.4 Objectives of the Study**

The primary objective of the project is to predict a start-ups operation status based on its financial information. The dataset is extremely biased, with more companies in ‘operating’ status. Therefore, the predictor should be accurate as well as precise and not using over/under sampling techniques.

The other objectives are:

- Clean the data with minimum data lost
- Apply feature engineering and keep only those features that suits best for the prediction
- Data modelling using different predictors and its validation
- Create an app using the final model and its deployment

## *Chapter – II*

---

---

# *Research Methodology*

---

---

## 2.1 Source of Data

Secondary source data is utilized for the completion of this project. The data is collected from Crunchbase, a platform that give insights into the business information about the public and private companies all over the world [<https://www.crunchbase.com/> ].

The scrapped data is pre-processed and feature engineering techniques are applied to the cleaned data before modelling and prediction.

## 2.2 Summary of Data

The data is scrapped from Crunchbase. It has 44 columns and 196553 rows. Among the total columns below are the major columns available in the dataset:

<b>name</b>	name of the company
<b>permalink</b>	website link
<b>category_code</b>	the category under which the company comes
<b>status</b>	status of the company, if it is ‘operating’, ‘acquired’, ‘IPO’ or ‘closed’
<b>founded_at</b>	data of foundation
<b>closed_at</b>	date of closure, if applicable
<b>first_investment_at</b>	date the company received its 1st investment
<b>last_investment_at</b>	date the company received its latest investment
<b>investment_rounds</b>	total number of investments

<b>first_funding_at</b>	date the company received the 1st funding
<b>last_funding_at</b>	date the company received the latest fundings
<b>funding_rounds</b>	total number of funding
<b>funding_total_usd</b>	total amount of funds received
<b>first_milestone_at</b>	date the company achieved its 1st milestone
<b>last_milestone_at</b>	date the company achieved its latest milestone
<b>milestones</b>	total number of milestones achieved
<b>lat</b>	latitude of the location of the company
<b>lng</b>	longitude of the location of the company
<b>ROI</b>	Rate of Investments

## 2.3 Statistical Tools Used

Python is the major tool used for the completion of the project. Excel is used to support the data visualization.

Below is the list of python libraries used in the analysis:

1. **Pandas** – open-source data analysis and manipulation tool, built on top of the Python programming language.
2. **NumPy** – used for working with arrays and mathematical functions
3. **Matplotlib** – used for creating static, animated, and interactive visualizations in Python
4. **Seaborn** – data visualization library based on matplotlib.
5. **Warnings** – used in situations where it is useful to alert the user of some condition in a program

6. **Datetime** – to work with dates and times
7. **Plotly** – is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases
8. **Sci-kit learn** – It provides efficient tools for machine learning and statistical modeling (sklearn.feature\_selection, sklearn.ensemble, sklearn.preprocessing, sklearn.model\_selection, sklearn.discriminant\_analysis)

## 2.4 Softwares Used

Jupyter is used for pre-processing, EDA, feature engineering and modelling. Deployment is done with the help of PyCharm, Streamlit, Heroku and GitHub platforms.

- **Jupyter notebook** is a browser-based open-source data science tool that supports Python.
- **PyCharm** is a dedicated IDE tool focused on providing a complete solution for creating full-fledged packages and software in Python, including classes and graphical user interfaces (GUIs).
- **Streamlit** is an open-source app framework for Machine Learning and Data Science teams.
- **Heroku** is a platform as a service (PaaS) that enables developers to build, run, and operate applications entirely in the cloud.
- **GitHub** is an Internet hosting service for software development



## **2.5 Scope of the Study**

The project aims to predict the status of a start up by analysing its financial parameters. Different start-ups from all over the world is considered for the project. The data is scrapped from Crunchbase database and it provides information about different start-ups from different geographical locations.

## **2.6 Significance of the Study**

An investor, before investing in any start-up would like to know what would be the status of that company after few years, given its financial information. Based on this forecasting they plan to invest their money in any start-ups.

It is at this point the significance of the status prediction application comes into picture. The application takes financial information of the company as input and the output is the predicted status of the company.

## **2.6 Limitations**

The data taken for the analysis is biased as more than 80% of the companies are in ‘operating’ status. For an extremely biased data, a constant predictor can give high accuracy but at the cost of lower precision.

QDA and Random Forest techniques are used to overcome this limitation, to an extent.

## *Chapter - III*

---

---

### *Company & Industry Profile*

---

---

### 3.1 Company Profile



Link – <https://technocolabs.com/>

Technocolabs was founded in 2019 by a Team of Non-Profit Organization in Indore, Madhya Pradesh. The primary area of focus of the company is Machine Learning, Data Science and Artificial Intelligence based product development. The Chief Executive Officer of Technocolabs is Mr. Yasin Shah.

The technologies that they utilise are Python, Java, Ruby, Ruby on Rails, microservices architecture, Elixir, REST API, WebSockets, JavaScript, Vue.js, React, React Native, Redis, ELK Stack, RabbitMQ, PostgreSQL, MongoDB, nginx, Amazon Web Services: EC2, S3, CloudFront. They cooperate mainly with E-commerce, Marketplace, FinTech, SaaS and AdTech clients.

**Mission** – “Our mission is to enhance business growth of our customers with creative design, development and to deliver market defining high quality solutions that create value and reliable competitive advantage to customers around the globe.”

**Vision** – “We believe that we are on the face of the earth to make great products and that's not changing. We are constantly focusing on innovating. We believe in the simple not the complex. We believe that we need to own and control the primary technologies behind the products that we make and participate only in markets where we can make a significant contribution.”

**Plan** – “Our plan is to setup requirements according to our clients and customers satisfactions and proper understanding.”

**Care** – “Our values define us as a company. They are a source of inspiration as we lead the way to a brighter future for our company and all who depend on it. They support our mission of making the lives better for each and every client and patient we care for.”

**Services Offered** – Machine Learning, Computer Vision, Mobile App Development, Voice Enabled Skill, Web Application Development, Big Data and Data Science.

**Approved by** – Ministry of Corporate Affairs, Govt. of India

### **3.2 Industry Profile**

The global business software and services market size was valued at USD 429.59 billion in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 11.7% from 2022 to 2030. The growing volume of enterprise data and increased automation of business processes across industries such as retail, manufacturing, and healthcare are driving the market growth. Moreover, the rapid deployment of enterprise software and services across IT infrastructure to improve decision-making, reduce inventory cost, and enhance profitability is also contributing to market growth.

Business for development, maintenance and publication of software using different business models like license / maintenance based or Cloud based are the ones that comes under the software industry. It also includes training, documentation, consulting and data recovery.

Business expansion initiatives by several companies across the world are expected to fuel market growth. The rapidly increasing use of cloud platforms, owing to benefits

such as flexibility, cost-effectiveness, and mobility, has triggered the demand for cloud-based software solutions and services among small and medium-sized businesses. Also, the market is expected to benefit from the rising use of innovative technologies such as blockchain, hybrid architecture, artificial intelligence, and machine learning over the forecast period.

The COVID-19 pandemic had a favourable impact on the business software and services market. According to an NTT Ltd. report commissioned by International Data Group, Inc. (IDG), the institutionalization of the work-from-home model amid local and worldwide quarantines has boosted the demand for value-added services for mitigating security concerns. Moreover, economic uncertainties caused by the pandemic have encouraged several vendors to focus on customer service-driven methods, including proactive support in customers' digital journeys.

## *Chapter – IV*

---

---

# *Data Analysis and Findings*

---

---

## 4.1 Data Pre-Processing

Data pre-processing can be referred to manipulation or dropping of data before it is used. It ensures the performance of the data in further steps of analysis.

Pandas, NumPy, Matplotlib, Seaborn, Warnings and Datetime libraries are used for data pre-processing.

The data is scrapped from Crunchbase. It has 44 columns and 196553 rows.

Below are the steps performed for pre-processing:

### A. Delete irrelevant and redundant information

- a. Delete 'region', 'city', 'state\_code' as they provide too much of granularity.
- b. Delete 'id', 'Unnamed: 0.1', 'entity\_type', 'entity\_id', 'parent\_id', 'created\_by', 'created\_at', 'updated\_at' as they are redundant.
- c. Delete 'domain', 'homepage\_url', 'twitter\_username', 'logo\_url', 'logo\_width', 'logo\_height', 'short\_description', 'description', 'overview', 'tag\_list', 'name', 'normalized\_name', 'permalink', 'invested\_companies' as they are irrelevant features.
- d. Delete duplicate values if any. e. Delete those which has more than 98% of null values.

### B. Remove noise or unreliable data (missing values and outliers)

- a. Delete instances with missing values for 'status', 'country\_code', 'category\_code' and 'founded\_at'.
- b. Delete outliers for 'funding\_total\_usd' and 'funding\_rounds'.
- c. Delete contradictory (mutually opposed or inconsistent data).

**C. Changes in original data**

- a. Convert `founded_at`, `closed_at`, `first_funded_at`, `last_funding_at`, `first_milestone_at`, `last_milestone_at` to years.
- b. Generalize the categorical data i.e., `category_code`, `status` and `category_code` - using One-Hot Encoding

**D. Creating new variables:**

- a. Create new feature `isClosed` from `closed_at` and `status`. `'isClosed' = 0` if `'status' = 'operating' or 'ipo'` `'isClosed' = 1` if `'status' = 'acquired' or 'closed'`
- b. Create new feature `'active_days'`  
`'active_days'` calculated using `'closed_at'`, `'founded_at'`, and `'status'` to calculate the age of the company

**E. Other transformations:**

- a. Delete `'closed_at'` column
- b. Fill null values of the numerical with the mean values of each column
- c. Drop columns with null values, i.e., `'first_investment_at'`, `'last_investment_at'`, and `'state_code'`

Save the dataset to a new file.

**Output** – After the pre-processing the stage the resultant dataset contained 9808 rows and 164 columns. The count of rows decreased due to the removal of null and duplicate values. The increase in number of columns is due the application of One-Hot encoding to generalize the categorical data.



## 4.2 Exploratory Data Analysis

Exploratory Data Analysis or EDA refers to the process of exploration of data to discover the patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Pandas, NumPy, Matplotlib, Seaborn, Plotly and Warnings libraries are used for performing EDA.

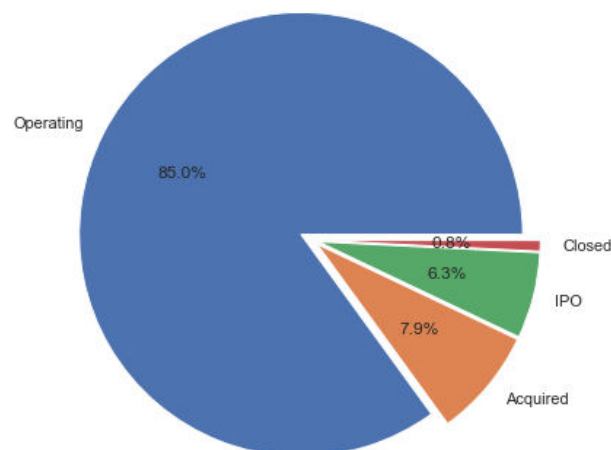
Pie chart, scatter plot, column graphs, etc are used to plot the different frequency distributions. Using 'lat' and 'lng' columns an attempt to plot the locations of the start-ups is also made.

### 4.2.1. The Status

The main objective of the project is to predict the status of the company, if it is operating, acquired, IPO or closed. Hence, it is important to analyze the target column.

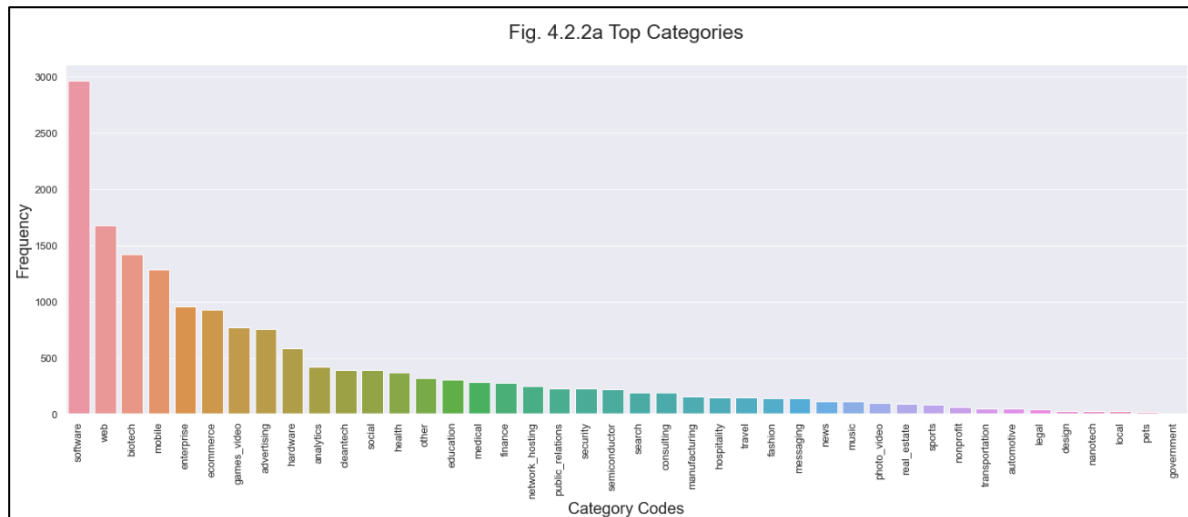
Below graph shows the distribution of the data over the four statuses of the companies:

Fig. 4.2.1 Distribution of Companies based on its Functional Status

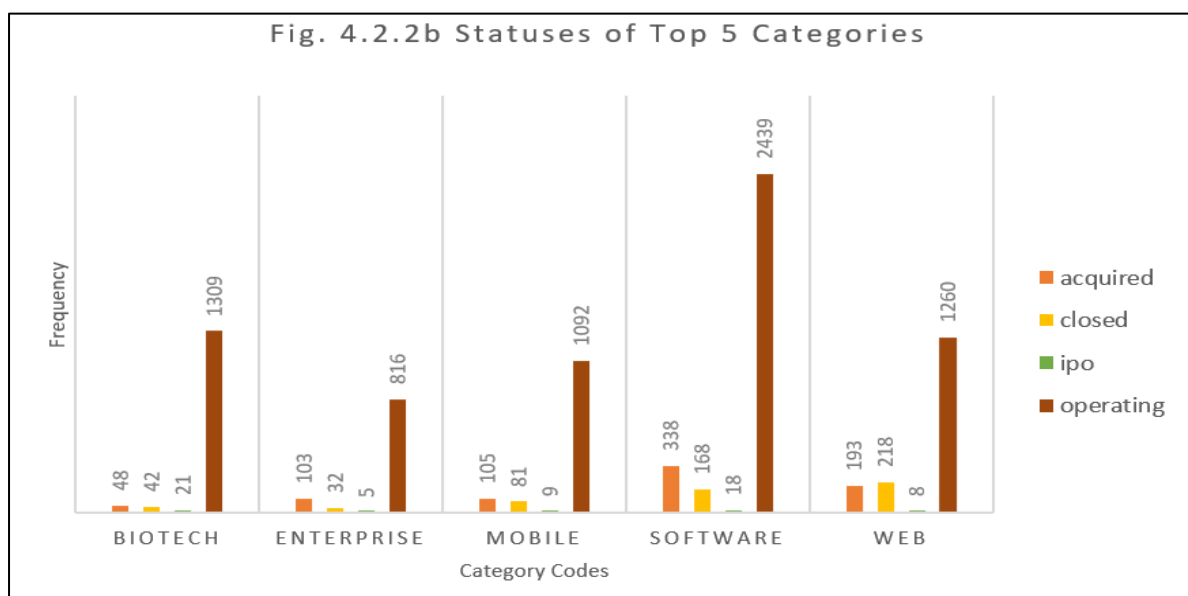


From the above pie chart, we see that 85% of the companies in the dataset are in ‘operating’ status. The dataset is biased, as it contains greater number of companies in ‘operating’ status, when compared to other statuses.

#### 4.2.2. Top Categories

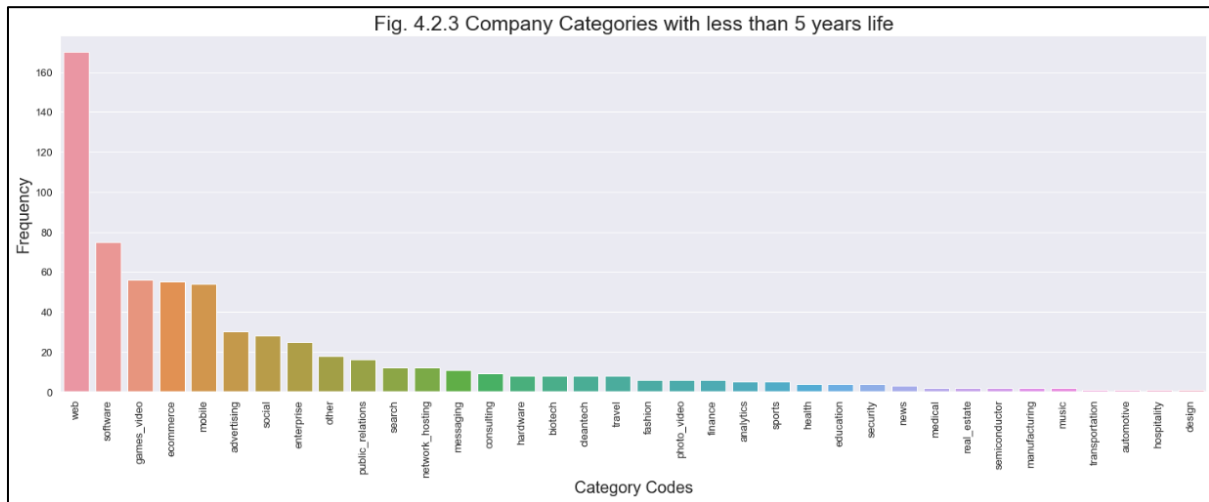


From fig 4.2.2a, we can observe that 'software', 'web', 'biotech', etc are the top categories, under which there are more companies. Fig. 5.2.2b shows the distribution of statuses over the top 5 categories.



### 4.2.3 Life < 5 Years

Fig 4.2.3 shows the companies under different categories which had only less than 5 years of life.



### 4.2.4 'Operating' status

Fig 4.2.4a Categories – Operating

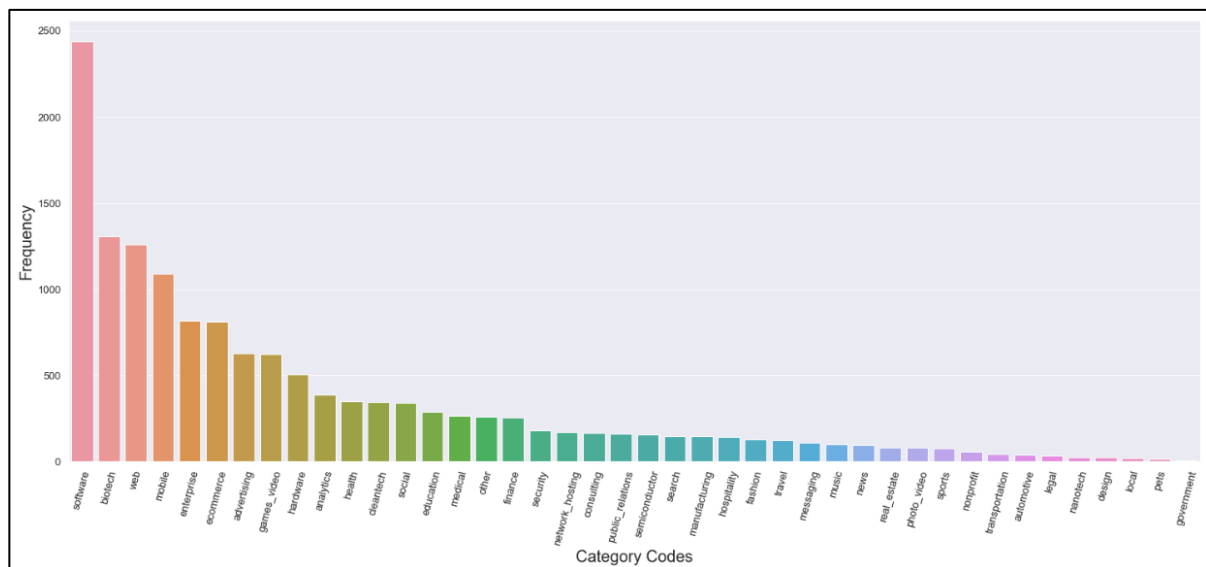


Fig. 4.2.4a shows the different categories of companies that are in operating status.

Fig 4.2.4b Origination Year – Operating

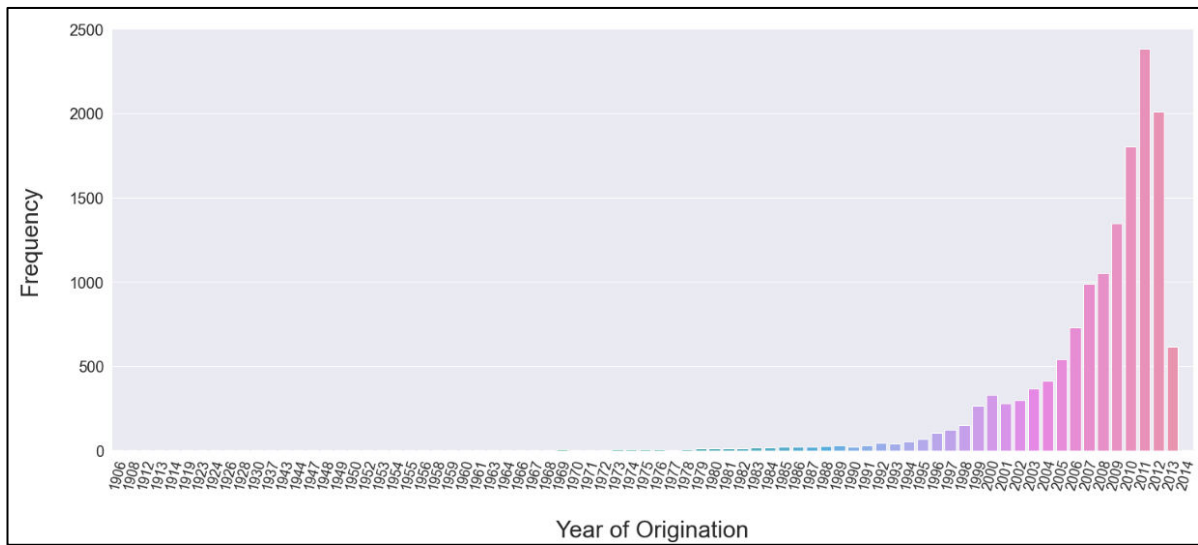


Fig. 4.2.4b shows the count of operating companies originated in different years.

## 4.2.5 ‘IPO’ status

Fig 4.2.5a Categories - IPO

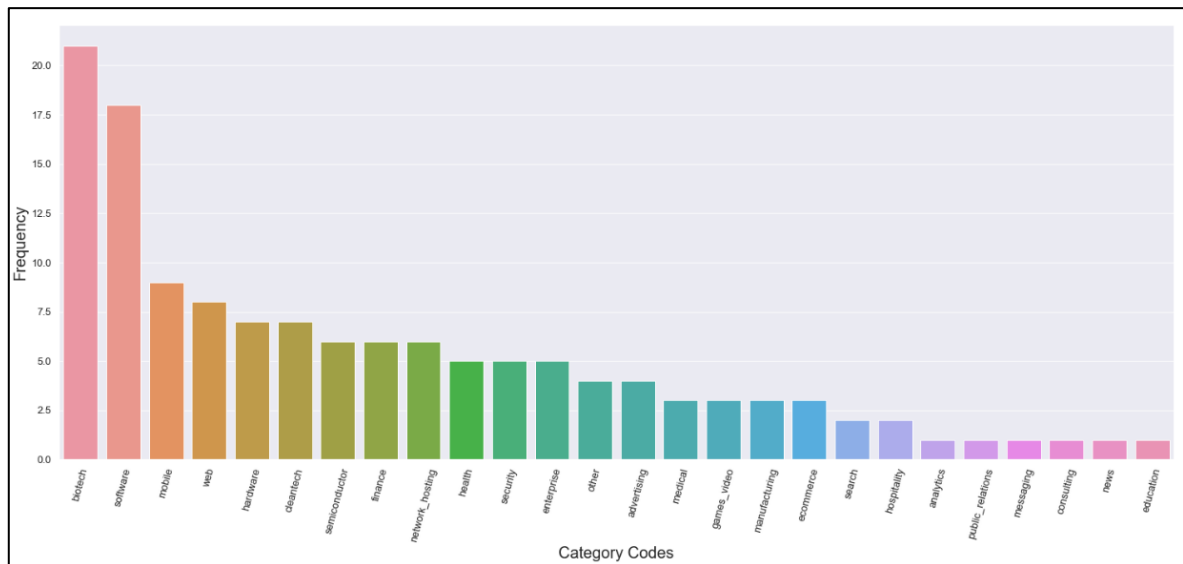


Fig. 4.2.5a shows the different categories of companies that are in IPO status.

Fig 4.2.5b Origination Year – IPO

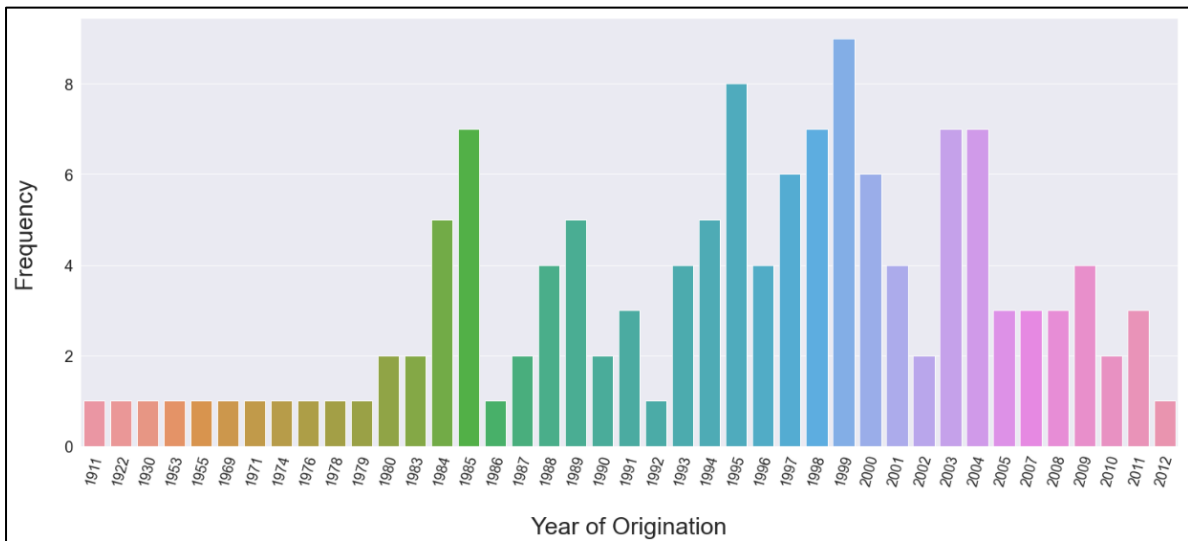


Fig. 4.2.5b shows the count of IPO companies originated in different years.

## 4.2.6 ‘Acquired’ status

5.2.6a Categories – Acquired

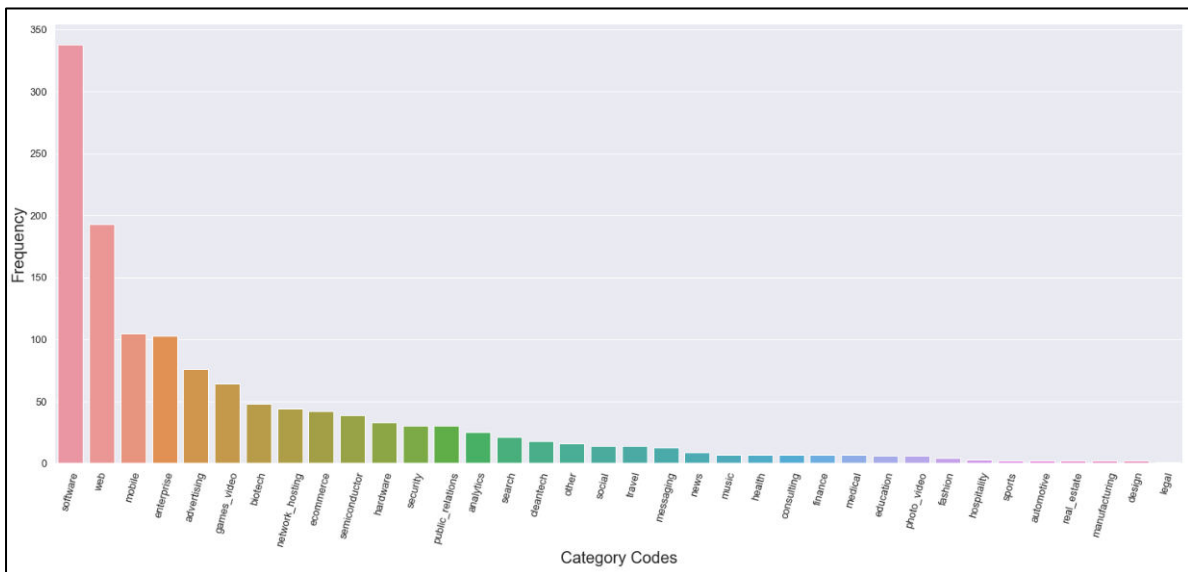


Fig. 4.2.5a shows the different categories of companies that are in IPO status.

Fig 4.2.6b Origination Year – Acquired

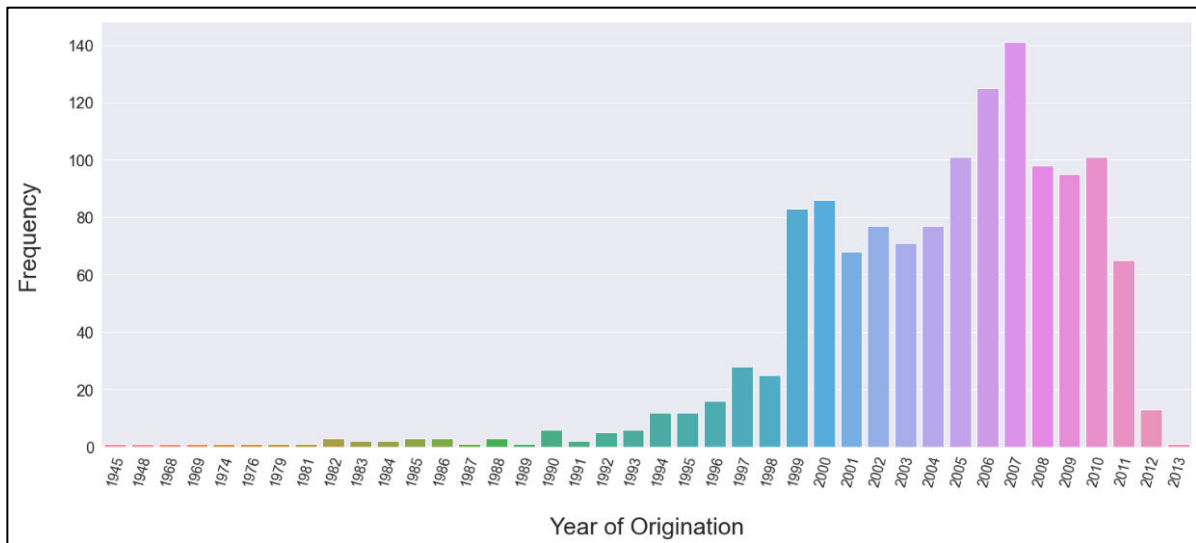


Fig. 4.2.6b shows the count of acquired companies originated in different years.

Fig 4.2.6c Closing Year – Acquired

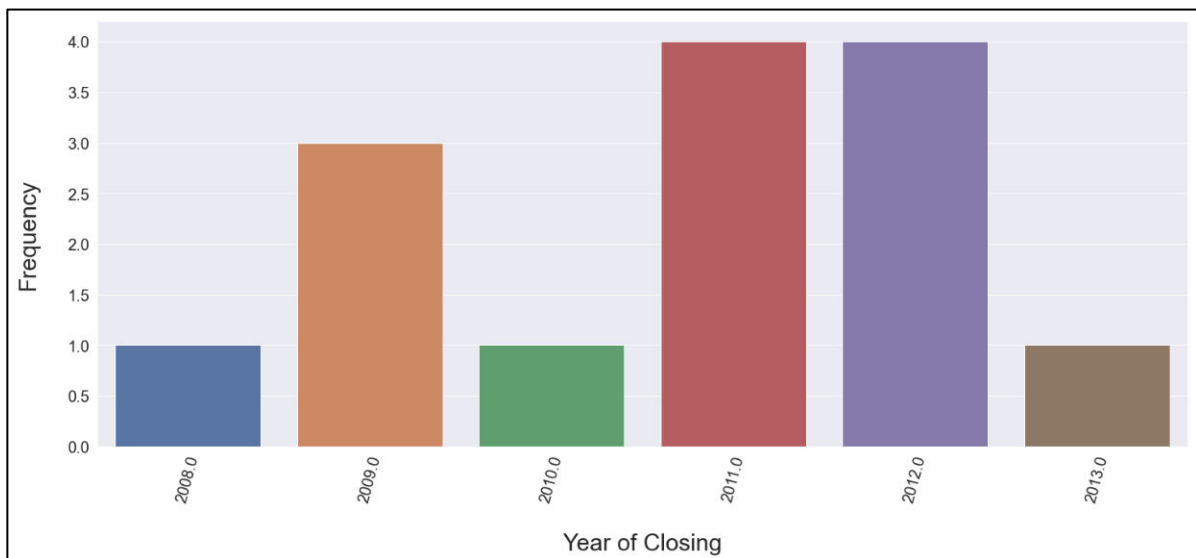


Fig. 4.2.6c shows the count of acquired companies closed in different years.

## 4.2.7 'Closed' status

Fig 4.2.7a Categories - Closed

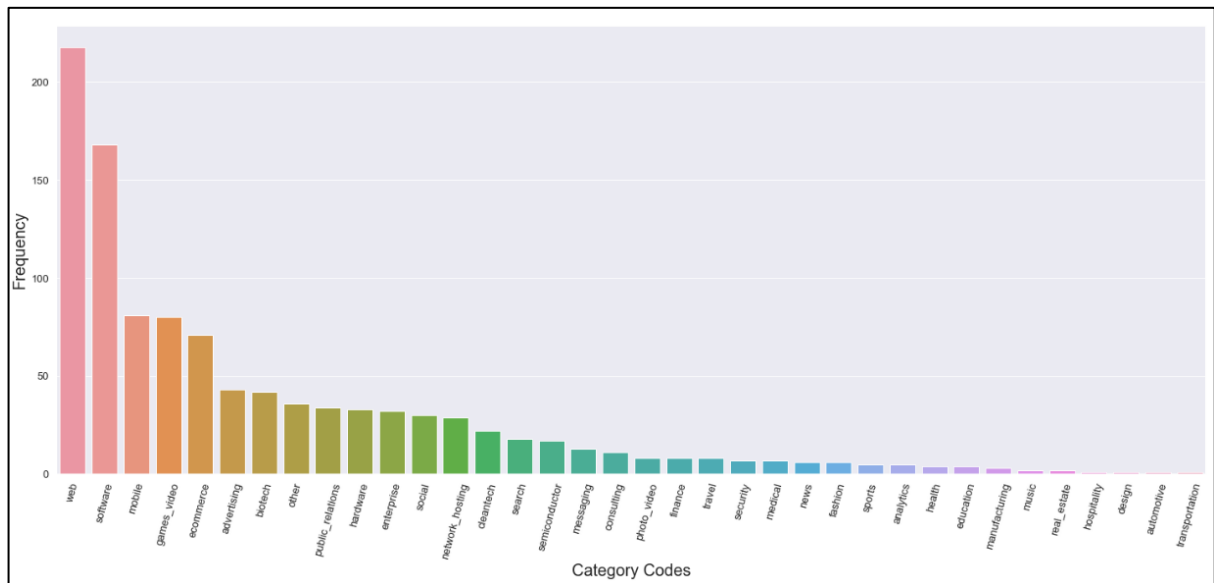


Fig. 4.2.7a shows the different categories of companies that are in closed status.

Fig 4.2.7b Origination Year – Closed

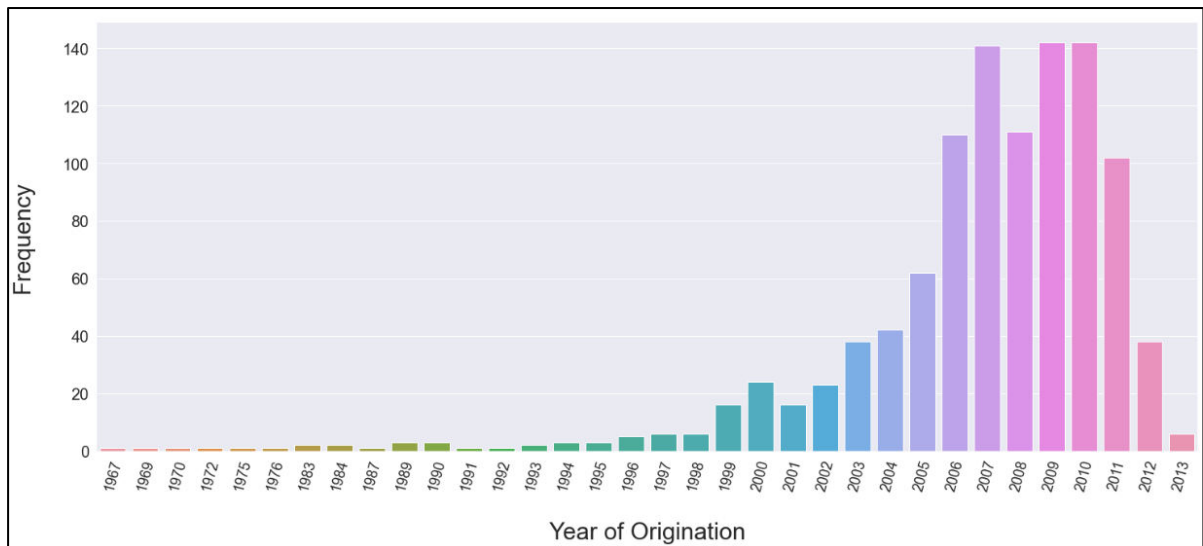


Fig. 4.2.7b shows the count of closed companies originated in different years.

Fig 4.2.7c Closing Year – Closed

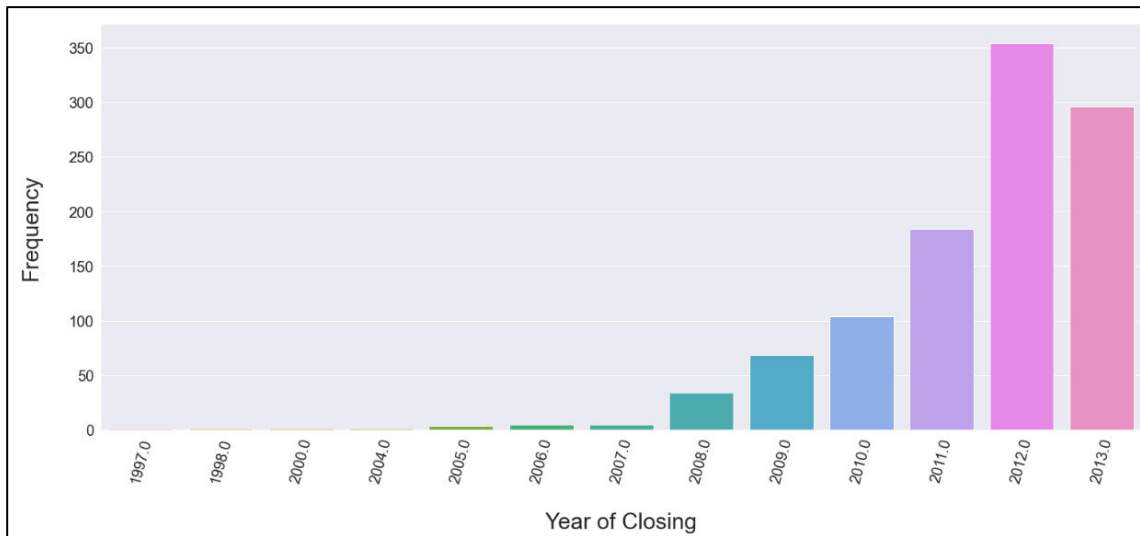
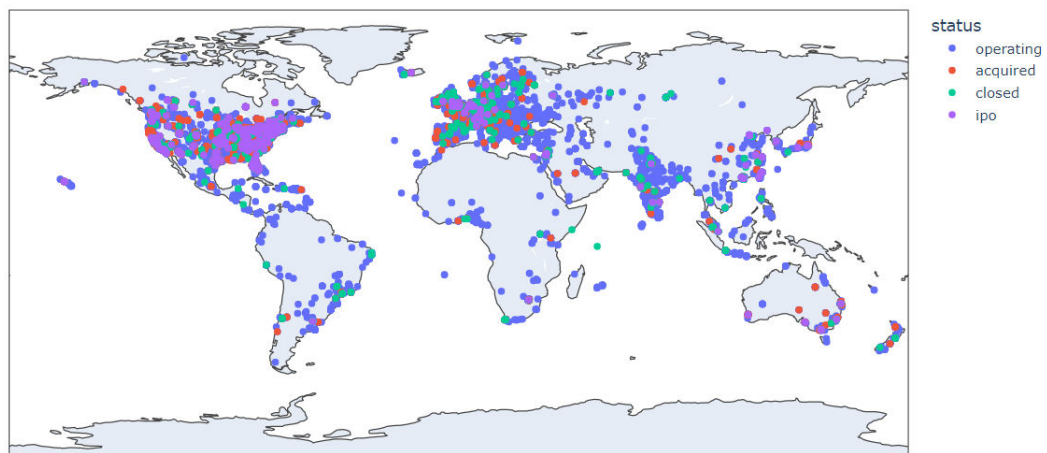


Fig. 4.2.7c shows the count of closed companies closed in different years.

### 4.2.3 Geographical Location vs Status

Fig. 4.2.3 shows the different geographical locations of the companies taken into consideration for the project.

Fig. 4.2.3 Startups Locations vs Status



## 4.3 Feature Engineering

Feature engineering is the technique by which features are extracted from the raw data to be used for the model creation.



Pandas, NumPy, Matplotlib, Seaborn and Sci-kit learn (sklearn.feature\_selection, sklearn.ensemble, sklearn.preprocessing) are used for applying feature engineering techniques.

Below are the feature engineering techniques performed:

1. **Factorize** – Using the .factorize() method to transform the dataset into numerical type to find the correlations between them.

*Output* – The status column which was initially a character column is now a numerical column after the application of factorize function. The values are updated as below:

0 – Acquired

1 – Operating

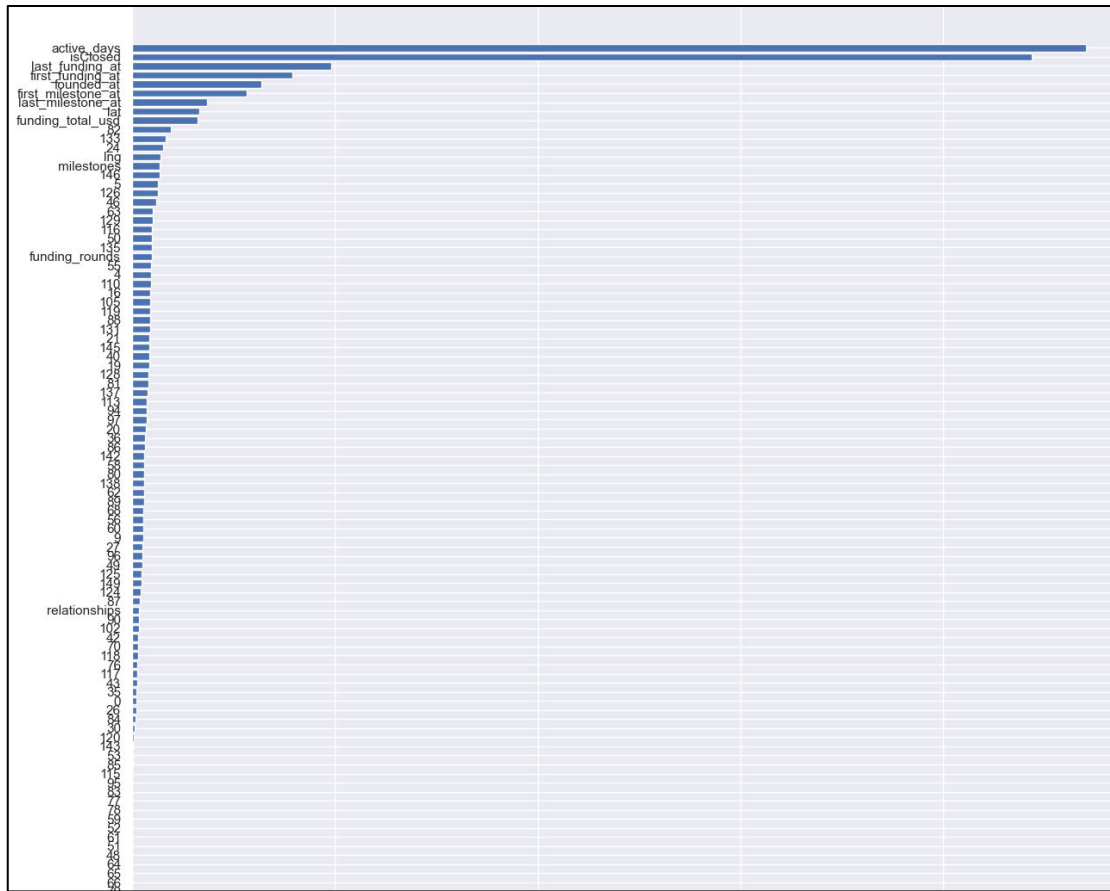
2 – Closed

3 – IPO

2. **Mutual Information** – Using MI to detect the most important features in the data. It is more efficient than the correlation as it doesn't assume a linear relationship, instead it measures the level of uncertainty between two features.

*Output –*

Fig. 4.3.1 Mutual Information Scores



From fig. 4.3.1, it can be observed that 'first\_funding\_at' and 'last\_funding\_at' have the highest scores. Note that 'active\_days' and 'isClosed' are the calculated columns while 'first\_funding\_at' and 'last\_funding\_at' were present in the initial scrapped data.

### 3. Univariate Analysis

Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes each variable on its own.

- Selection with SelectKBest - select the independent features which have the strongest relationship with the dependent feature

Feature with the highest score will be more related to the dependent feature.

**Output** – 'founded\_at', 'first\_funding\_at', 'last\_funding\_at', 'isClosed' and 'active\_days' are the top 5 columns resulted after applying SelectKBest method.

- Selection ExtraTreesClassifier – Helps to give the importance of each independent feature with a dependent feature

Higher score means it's more important/relevant towards output variable

**Output** – An output (fig. 4.3.2) similar to the SelectKBest is resulted after using ExtraTreesClassifier method also.

Fig. 4.3.2 – ExtraTreesClassifier Scores

	0
<b>isClosed</b>	0.226110
<b>active_days</b>	0.166200
<b>first_funding_at</b>	0.016067
<b>last_funding_at</b>	0.013519
<b>founded_at</b>	0.011761
<b>first_milestone_at</b>	0.008026
<b>last_milestone_at</b>	0.006100
<b>funding_total_usd</b>	0.004456
<b>milestones</b>	0.002945

'founded\_at', 'first\_funding\_at', 'last\_funding\_at', 'isClosed',  
'funding\_rounds', 'funding\_total\_usd', 'first\_milestone\_at', 'status',

'last\_milestone\_at' and 'milestones' are the columns selected for the further procedures.

**4. Data Standardization** – It a data processing workflow that converts the structure of different datasets into one common format of data.

It is used when all features are having high values, not 0 and 1. The mean of the independent features is 0 and the standard deviation is 1.

*Output –*

Fig. 4.3.3 - Standardization

	founded_at	first_funding_at	last_funding_at	funding_rounds	funding_total_usd	first_milestone_at	last_milestone_at	milestones	status	isClosed
0	-0.113605	-0.879874	-1.306764	-0.791392	0.001246	-0.753281	0.361501	1.733481	0	0.0
1	-0.799293	0.310334	0.454866	1.970875	0.773914	-0.084546	-0.357688	-0.735943	1	1.0
2	-0.799293	-2.863553	-3.508802	-0.791392	-0.714844	-1.087648	-1.436472	-0.735943	1	1.0
3	-0.627871	0.310334	0.014459	-0.791392	-0.526399	-0.084546	-0.357688	-0.735943	1	1.0
4	-0.285027	-1.276609	-1.747172	-0.791392	-0.375644	-0.084546	0.361501	0.498769	1	1.0
...	...	...	...	...	...	...	...	...	...	...
9803	0.057816	-0.879874	-1.306764	-0.791392	-0.732935	-0.753281	-0.717283	0.498769	2	0.0
9804	0.572082	0.310334	0.014459	-0.791392	-0.677155	0.249821	0.001906	-0.735943	1	1.0
9805	-0.285027	-0.879874	-0.866356	0.589742	-0.586702	-1.756383	-1.076877	0.498769	2	0.0
9806	-0.113605	-0.879874	-1.306764	-0.791392	-0.639466	0.918555	0.721096	-0.735943	1	1.0
9807	-0.113605	-0.879874	-1.306764	-0.791392	-0.680924	-1.422016	-1.076877	0.498769	1	1.0

9808 rows × 10 columns

**5. Save the dataset to a new file**

## *Chapter – V*

---

### *Model Building*

---

## 5.1 Model Building

A machine learning model is a file that has been trained to recognize certain types of patterns.

The data set is separated into training and testing data. An appropriate model is selected for the dataset and using the training data, an algorithm is provided to the model to learn the patterns in the data.

Sci-kit learn (sklearn.discriminant\_analysis, sklearn.ensemble, sklearn.model\_selection, sklearn.metrics) along with other packages such Pandas, NumPy, Matplotlib, Seaborn and Warnings are used for model building.

### 5.1.1 Quadratic Discriminant Analysis

QDA is a method designed to separate two or more classes based on a combination of features in a normal distribution, where it assumes each feature has its own covariance matrix.

It is more flexible with high variance data and is mainly used to classify between two classes only, Operating and Not Operating.

QDA uses a classifier with a quadratic decision boundary, where each class is fitted with a Gaussian density.

***Output*** – Accuracy Score for prediction = 0.835236541598695

Fig. 5.1.1 – QDA Prediction

	Original	Prediction	Error
9248	1	1	0
1463	1	0	1
4928	1	1	0
3057	1	1	0
2450	1	1	0
...	...	...	...
6967	1	1	0
1916	1	1	0
3079	1	1	0
5512	0	1	-1
5197	1	1	0
2452 rows × 3 columns			

### 5.1.2 Random Forest Classifier

Classification algorithm consisting of many individual decision trees that operate as an ensemble.

Each tree predicts a decision/class, and the decision with the most votes become the final prediction.

Low correlation between data features helps this ensemble reach better scores.

This classifier is used to classify between all 4 classes, Operating, IPO, Acquired, and Closed.

**Output** – Accuracy Score for prediction = 0.835236541598695

Fig. 5.1.2 – Random Forest Prediction

	Original	Prediction	Error
9248	1	1	0
1463	1	1	0
4928	1	1	0
3057	1	1	0
2450	1	1	0
...	...	...	...
6967	1	1	0
1916	1	1	0
3079	1	1	0
5512	2	2	0
5197	1	1	0
2452 rows × 3 columns			

## 5.2 Model Validation

Mean Absolute Error - the magnitude of difference between the prediction of an observation and the true value of that observation

*Output\_*–

QDA – 0.16476345840130505

RF – 0.20187601957585644

Cross Validation - tells how well a classifier generalizes, specifically the range of expected errors of the classifier.

*Output\_*–

QDA - array([0.82608696, 0.82121006, 0.84296397, 0.82596873, 0.82256968])



Confusion Metrix - the visual representation of the Actual vs Predicted values

*Output\_*–

Fig. 5.2.1 – Confusion Matrix for QDA

```
Confusion Metrix for Prediction Data :  
[[ 119 261]  
 [ 143 1929]]
```

Fig. 5.2.2 – Confusion Matrix for RF

```
Confusion Metrix for Prediction Data :  
[[ 29 199 5 0]  
 [ 53 1949 42 1]  
 [ 7 128 12 0]  
 [ 0 20 7 0]]
```

Classification Report - used to measure the quality of predictions from a classification algorithm.

*Output\_*–

Fig 5.2.3 – Classification Report for QDA

```
Classification Report for Prediction Data :  
  
              precision    recall  f1-score   support  
  
    0               0.45        0.31        0.37         380  
    1               0.88        0.93        0.91        2072  
  
 accuracy               0.84         2452  
 macro avg              0.67        0.62        0.64         2452  
 weighted avg           0.81        0.84        0.82         2452
```

Fig. 5.2.4 – Classification Report for RF

Classification Report for Prediction Data :				
	precision	recall	f1-score	support
0	0.33	0.12	0.18	233
1	0.85	0.95	0.90	2045
2	0.18	0.08	0.11	147
3	0.00	0.00	0.00	27
accuracy			0.81	2452
macro avg	0.34	0.29	0.30	2452
weighted avg	0.75	0.81	0.77	2452

Accuracy Score - number of correctly classified prediction to the total number of predictions.

**Output –**

Accuracy Score for QDA = 0.835236541598695

Accuracy Score for RF = 0.835236541598695

### 5.3 Predictor Function

User defined function that combines qda\_model results and rf\_model results to give the final prediction.

Fig. 5.3.1 – Predictor Function Prediction

	Original	Prediction	Error
0	1	1	0
1	1	3	-2
2	1	1	0
3	1	1	0
4	1	1	0

## *Chapter – VI*

---

# *Model Deployment*

---

## 6.1 Model Deployment

Model deployment is the process of putting machine learning models into production.

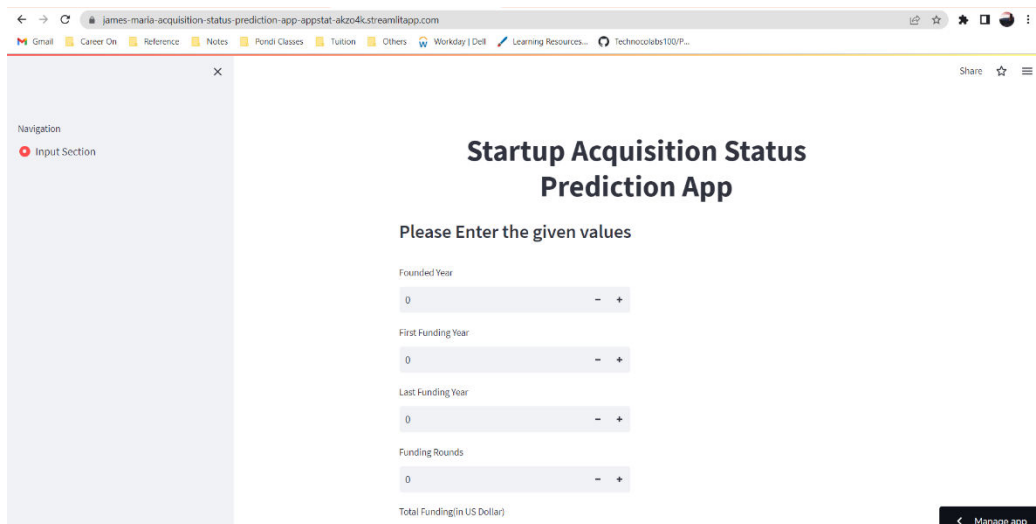
It allows the users or developers to access the model's predictions, so that they can make business decisions based on data, interact with their application, etc.

Streamlit is used for deployment along with other libraries such as Pandas, NumPy and Sci-kit learn (sklearn.discriminant\_analysis, sklearn.ensemble, sklearn.model\_selection, sklearn.preprocessing)

The application is named as AppStat (alias stat-pred). It is built using Streamlit and deployed using Streamlit as well as Heroku.

App deployment using Streamlit Link – <https://james-maria-acquisition-status-prediction-app-appstat-akzo4k.streamlitapp.com/>

**Output -**



The screenshot shows a web browser window displaying the 'Startup Acquisition Status Prediction App'. The browser's address bar shows the URL: [james-maria-acquisition-status-prediction-app-appstat-akzo4k.streamlitapp.com](https://james-maria-acquisition-status-prediction-app-appstat-akzo4k.streamlitapp.com). The app's interface includes a navigation sidebar on the left with a red dot next to 'Input Section'. The main content area is titled 'Startup Acquisition Status Prediction App' and contains the instruction 'Please Enter the given values'. Below this, there are five input fields, each with a '0' value and minus/plus buttons: 'Founded Year', 'First Funding Year', 'Last Funding Year', 'Funding Rounds', and 'Total Funding(In US Dollar)'. A 'Manage app' button is located in the bottom right corner of the app's interface.

Total Funding(in US Dollar)

0.0000 - +

First Milestone Year

0 - +

Last Milestone Year

0 - +

Total Milestones

0 - +

← Manage app

App deployed via Heroku Link – <https://stat-pred.herokuapp.com/>

Navigation

Input Section

Startup Acquisition Status Prediction App

Please Enter the given values

Founded Year

0 - +

First Funding Year

0 - +

Last Funding Year

0 - +

Funding Rounds

0 - +

Total Funding(in US Dollar)

## *Chapter - VII*

---

### *Suggestions & Conclusion*

---

## **7.1 Suggestions**

QDA fails to increase precision because there are not enough minority points to accurately determine the distribution. At this point, RF model increases the precision but overfits the training data. Combining both the models, a higher precision model helps to increase the precision without decreasing the accuracy.

Using more effective over / under – sampling techniques, the model can be tuned further to increase the proficiency. Other model combination can also be tried to verify if it increases the accuracy.

## **7.2 Conclusion**

Several world-famous companies have started its journey as a start-up. Several start-ups have taken its own position in the global economy. Investors who are willing to invest their money in a start-up is likely to know the future status of the venture. The project is all about the prediction app which is based on the prediction of the start-ups acquisition status.

Beginning with a scrapped data from Crunchbase database, the project moved ahead with the data cleaning, applying feature engineering techniques to get accurate data features for prediction, model building and prediction. The model predicted the status with an accuracy of 78.5 %. The model was then deployed using Streamlit and Heroku. The report clearly mentions about the different tools used for the completion of the project and the step-by-step process followed.

The application can be used for predicting the status of a start-up, by giving few financial information about the company. In future, the model can be further improvised by providing the details of emerging start-ups.

## REFERENCES

1. <https://github.com/Technocolabs100/Predictive-Analysis-on-Startup-Acquisition-Status--DA15072022>
2. <https://github.com/James-Maria/Acquisition-Status-Prediction-App>
3. <https://technocolabs-internship.gitbook.io/internship-prerequisites-learning-resources/>
4. <https://www.analyticsvidhya.com/blog/>
5. <https://www.geeksforgeeks.org/>