Startup’s Acquisition Status prediction

Prepared by: Mostafa Samy Ibrahem

Abstract

Start-ups are companies that make items that wander to a zone or showcase in ways that haven't been done some time recently. This makes start-ups unsafe and unusual as a modern item or benefit may not work among its clear clients and may require consistent alterations some time recently it gets product/market fit. Eventually, a start-up could be a high-risk company that's within the to begin with arrangement of operations and commonly related to innovation as an item or a benefit. Startups are important and the motor for the economy of the countries.

Over the past decade, many countries have seen an exponential growth in start-up arrangements. In this way, it appears a significant challenge understanding what makes this type of high-risk wanders effective and as such, appealing to financial specialists and business visionaries. Success for a start-up is characterized here as the occasion that gives an expansive whole of cash to the company's founders, speculators and early representatives.

The capacity to foresee victory is an invaluable competitive advantage for venture capitals on the rummage around for ventures since first-rate targets are those who have the potential for developing quickly before long, which eventually, permits financial specialists to be one step ahead of competition.

We investigated the world's biggest organized database for start-ups – given by the website CrunchBase.com, with the objective of building a prescient show, through directed learning, to accurately classify which start-ups are fruitful and which aren't. Most of the considers with respect to the prediction of forms of M&A or an elective definition of a company's victory tend to center on traditional administration measurements given by budgetary reports and in this way employing a moo number of observations compared with the display ponder. As advances of information evolve it became possible to attain exceedingly solid comes about in information examination by controlling it with complex machine learning calculations or information mining procedures to characterize highlights and characterize strong models.

Keywords

Start-up, Mergers and Acquisitions (M&A), IPO, data analysis, machine learning, true positive rate, false positive rate.

Agenda

Introduction

"A startup or start-up is a company or project undertaken by an entrepreneur to seek, develop, and validate a scalable business model."

Start-ups are booming all over as more colleges, governments and private companies contribute and stimulate individuals to seek after their thoughts all through these wanders. Companies are raising millions with ease and accomplishing unicorn status (i.e., a one-billion-dollar valuation) in a matter of a long time. Slack, a informing app, accomplished it after working for 1.25 a long time (Kim, 2015). Illustrations like Uber and Airbnb are changing social orders in such impactful ways that control had to be made to keep pace with a modern reality. Start-ups are having such affect that, eventually it gets to be each investor's ambition to be portion of a expansive securing such as Facebook obtaining WhatsApp (another messaging app) for nineteen billion dollars which permitted Sequoia (a Wander Capital finance) to have a 50x return on speculation (Neal, 2014). But there's a capture, start-ups are companies with an assessed 90% probability of disappointment, which suggests a parcel of speculations without legitimate returns (Patel, 2015).

Anticipating the victory of a start-up is commonly characterized as two-way procedure that creates a large amount of cash to its originators, speculators and to begin with workers, as a company can either have an IPO (Initial Open Advertising) by attending to a open stock advertise (i.e. Facebook going open, allowing everyone to contribute within the company by buying offers being sold by its insiders within the U.S stock market) or, be procured by or consolidated (M&A) with another company (i.e. Microsoft procuring LinkedIn for $26B) where those who have already contributed get quick cash in return for their shares. This handle is regularly designated as an exit technique (Guo, Lou, & Pérez-Castrillo, 2015). This study will in this manner, consider both an IPO (Beginning Open Advertising) and a prepare of M&A (Mergers & Acquisitions) as the basic occasions that classify a start-up as effective.

With a center on how a start-up or an speculator seem investigate all this information for distant better;a much better;a higher;a stronger;an improved">a higher decision making in speculation methodology and money related pick up, the study extreme, by applying information mining and machine learning procedures, to make a prescient show that has as the subordinate variable a label to classify whether a start-up is (as of now) fruitful or not.

To generate the predictive model, three supervised machine learning algorithms were tested: Random Forest, Xgboost, VotingClassifier.

Objectives

The present work has as the main objective, the development of a predictive model to classify a startup whether it is (Operating, IPO, Acquired, or closed). The most important step to get into startip's acquisitions is knowing its financial statistics such as total funding dollars, funding dates, number of funding rounds and headquarter locations, also the IPO is an important feature to consider. An Initial Public Offering (IPO) refers to the process of offering shares of a private corporation to the public in a new stock issuance.
Previous studies tend to center primarily on administrative highlights and regularly outline the effect of monetary highlights related with funding (specially from Wander Capital reserves). It is planning to bridge this crevice by making funding-oriented features with great prescient effect in classifying effective companies. Also, there's room to make strides in the quality of the test by being more particular with companies or by way better treating the sum of meager information which is characteristic of this dataset.

Literature Review

In this section we will identify major themes related to the subject of this study.

Start-ups are companies that make items that wander to a zone or showcase in ways that haven't been done some time recently. This makes start-ups unsafe and erratic as an unused item or benefit may not work among its clear clients and may require steady alterations some time recently when it gets product/market fit. Eventually, a start-up could be a high-risk company that's within the to begin with arrangement of operations and commonly related to innovation as an item or a benefit.

-- **Startups can now be built for thousands rather than millions**: With a diminishing toll of product development by a calculation of 10 over the final decade (Hermann et al., 2015), it is presently cheaper than ever to construct innovation. Getting to instruments, open-source code, cheaper servers, and an ever-growing community of designers contributing to the dispersal of innovation around the globe allows everyone to construct, test and share its items. The most elevated representation of this reality is WhatsApp, which was bought for more than $19 billion dollars and had sixteen representatives.

-- **A higher resolution venture capital industry:** When Wander Capital (VC) 1 were required to spend millions of dollars on a venture, they had to form a little number of enormous wagers. Be that as it may, with the cost of innovation being less costly each year it has made an opportunity for other sorts of investors: blessed messengers, quickening agents and micro-VCs. These substances, with littler checks can make a whole lot of little wagers and offer assistance to a bigger number of start-ups. This life saver for small start-ups permits them to not search for extra exterior financing until afterward stages of improvement.

-- **Entrepreneurship developing its own management science**: When the primary wave of Information Era venture-backed software companies started within the 1970's, numerous business people connected its knowledge of Administration Science made by Henry Passage and his peers. However, particularly after the tremendous dotcom bubble burst within the last a long time of the nineties, numerous business visionaries started to realize start-ups were a distinctive reality with a distinctive run the show set. Forty a long time after the starting of the modern start-up period, Steve Clear with "The Four Steps to the Epiphany" and Eric Ries with "The Lean Start-up" laid the establishment for a unused administration science for start-ups, which has come to be known as the Incline Start-up Development. Additional time "entrepreneurs have gotten to be essentially way better at creating start-ups."

-- **Speed of consumer adoption of new technology**: As web got to be all around available, start- ups can be - from day one - what Steve Clear calls, a "micro-multinational" and individuals from all over the world can get to items from the inverse conclusion of the planet without any inconvenience (Blank, 2006). Google and Facebook demonstrate that the area is likely insignificant.

The success of a start-up is commonly defined as a two-way strategy as a company can either have an IPO (Public Initial Offering) by going to a public stock market, allowing its shareholders to sell shares to the public, or be acquired or merged (M&A) with another company where those who have previously invested receive immediate cash in return for their shares. This process is often designated as an exit strategy.

Methodology

The methodology here applied in the following steps:

1. Selection of the data to be processed by defining relevant tables from the entire structured CrunchBase Dataset
2. Preprocessing, by cleaning, selecting, and transforming data. At this stage we deal with missing values, dropping unnecessary columns, removing outliers, and other common problems.
3. An EDA is made before any further transformations
4. Feature engineering on the data.
5. Experimental Setup, where evaluation metrics are defined and several machine learning algorithms are chosen to test the data, whether it's Operating, IPO, Closed, Acquired.
6. Experiments results, when we draw conclusions.

Data Collection and Selection (CrunchBase Dataset from Kaggle)

The dataset used for this project is a kaggle dataset sourced from Crunchbase called: "Crunchbase 2013-Companies, Investors, etc."
There are nearly 196553 rows and 15 columns, each row of the dataset contains a startup's information.
The dataset labels show that the dataset is extremely biased. As shown in the table below. With the "Operating" class is over-presented.

IPO	Closed	Acquired	Operating
1.9%	3.1%	9.4%	85.6%

Name	Observations	Selected
category_code		
ROI		
first_investment_at		
last_investment_at		
first_funding_at		
created_by		
status		
founded_at		
closed_at		
country_code		
created_by		
state_code		

updated_by		
city		
region		
funding_rounds		
funding_total_usd		
id		
invested_companies		
Unnamed:0.1		
entity_type		
parent_id		
name		
normalized_name		
permalink		
domain		
first_milstone_at		
last_milestone_at		
milestones		
relationships		
lat		
homepage_url		
twitter_username		
logo_url		
logo_width_height		
short_description		
overview		
tag_list		

Note: to produce a dataset for the training task, only tables marked as selected will be used.

Data Preprocessing

Data preprocessing can often have a critical impact on the performance of the model.
The data preprocessing consists of 3 steps:

- Data Cleaning: Where we need to remove all redundant and irrelevant information from the database as well as duplicates, missing values and outliers.
- Data Selection: Which is defined as the data is selected from the dataset to the final dataset
- Data Transformation: The process of creating new variables or aggregating data from different tables to the final table.

Data Cleaning

The first step of preprocessing is treating all irrelevant and redundant information present in tables. The Crunchbase dataset has several columns(features) and instances(observation) whose context don't match the objective of predicting a startup's success.
Only a few duplicated instances were found in the database and all were removed.
The second step consists of eliminating noisy and unreliable data being the two most common cases of inconsistencies, Missing values and Outliers.
A missing value is a variable that has no data value stored in an observation. Missing value are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.
Outliers are excessively deviating value from the scale of the feature. An example of an outlier found in the dataset can be extremely high "total funding is USD" and "funding rounds"
Another type of inconsistent data can be misspellings or contradictory values, especially due to the crowdsourced nature of the in-use dataset. Wrong dates or presence of letters in numerical features are examples of frequently present inconsistent data.

Data Selection

Companies with founding dates between 1985 and 2014: Although some of this companies can not be considered as startup anymore due to their advanced age without a success event, they were at some point and some events as funding rounds who potentially brought them closer to success, so these companies will stay in the dataset.

Companies with category: to try to compare results with previous publications and being a category of a company something that influences, among other factors, its average age of success, we chose to only take companies with a category for further analysis. The company's category reflects both its industry as well as if it is a tech company or not.

Data Transformation

What is meant by data transformation is the application of mathematical modification to the value of a variable to extract more value than in its original state.
There is two ways to apply data transformation

1. Changes in the original data
2. New variables created

Changes in the original data

- Encoding category_code column
- For the country_code column there are 161 country code, One Hot Encoding will create a lot of columns so we will get the first 10 values ordered by repetition and call the rest “other”. Then apply One Hot Encoding.

New variables created

- “isClosed”: a binary feature tells us whether the startup is still running or closed. So it's 1 closed and 0 otherwise.
- “Active days”: determines how many days the startup has been running. It comes when we subtract the founded_at column from closed_at column(if the company has closed) or 2021 if the company is still running.

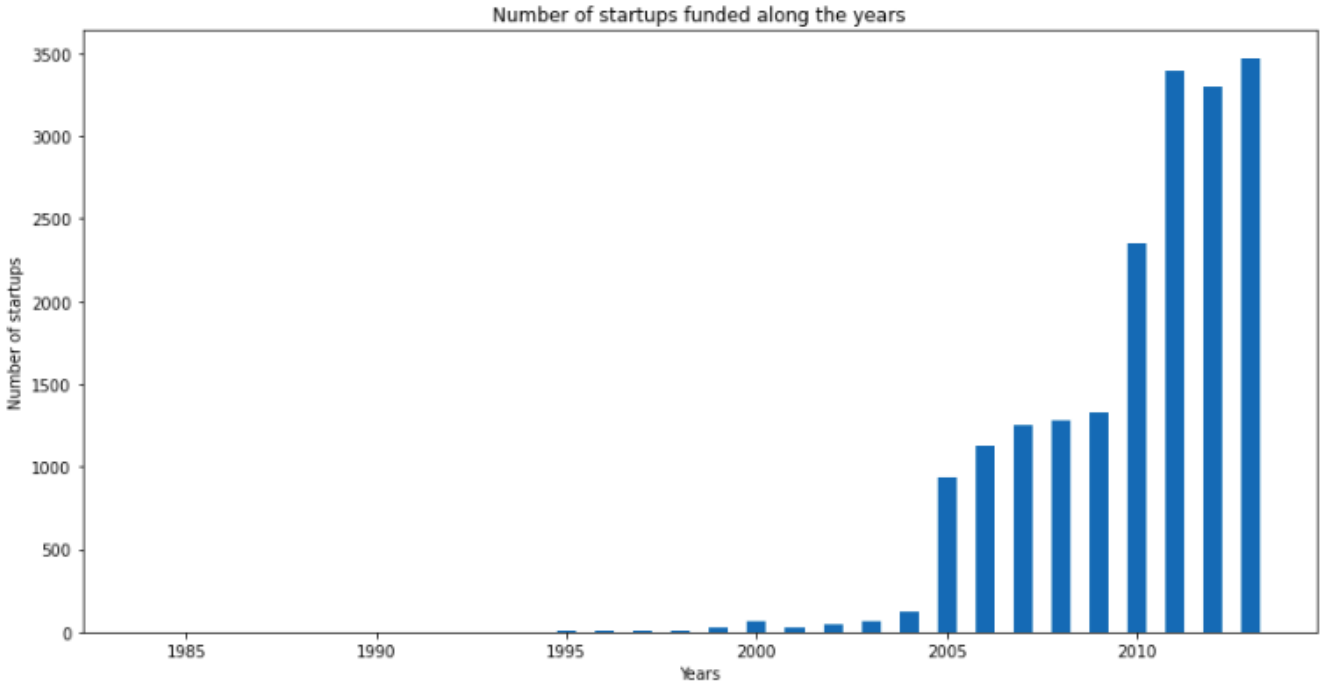
Exploratory Data Analysis

In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.
A statistical model can be used or not, but primarily EDA is for seeing what the data tells us beyond the formal modeling or hypothesis testing task.
For this project we have some research questions we want to know their answers through visualizations

- Research Question 1: How many startups have been released over the years? And which year has the most released number of startups?
- Research Question 2: What is the number of startups in every country?
- Research Question 3 : Which startup's category advances faster?
- Research Question 4: Who are the top 10 market leaders?
- Companies statues
- Status of startups in USA, UK, Canada, and China

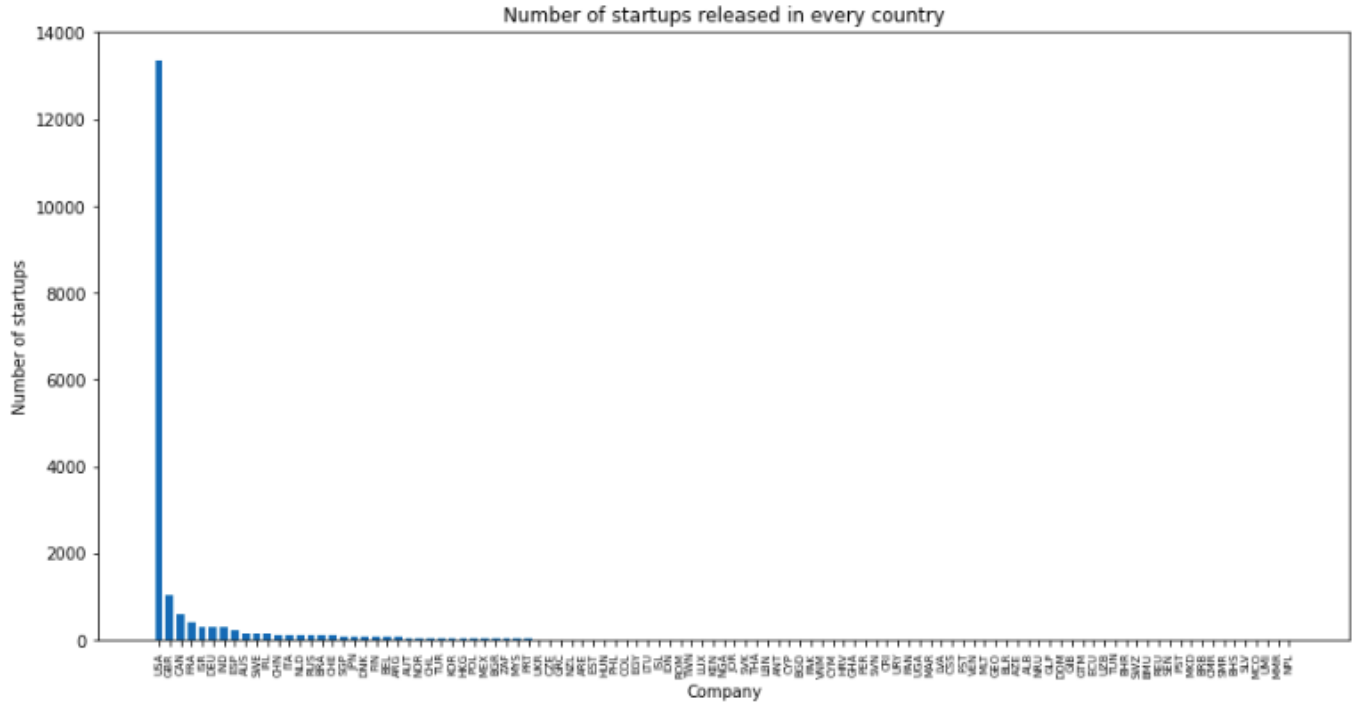
Research Question 1: How many startups have been released over the years? And which year has the most released number of startups?

As shown in the graph below, the distribution of the number of startups founded through the years, and its obvious that 2013 has the most released number of startups.



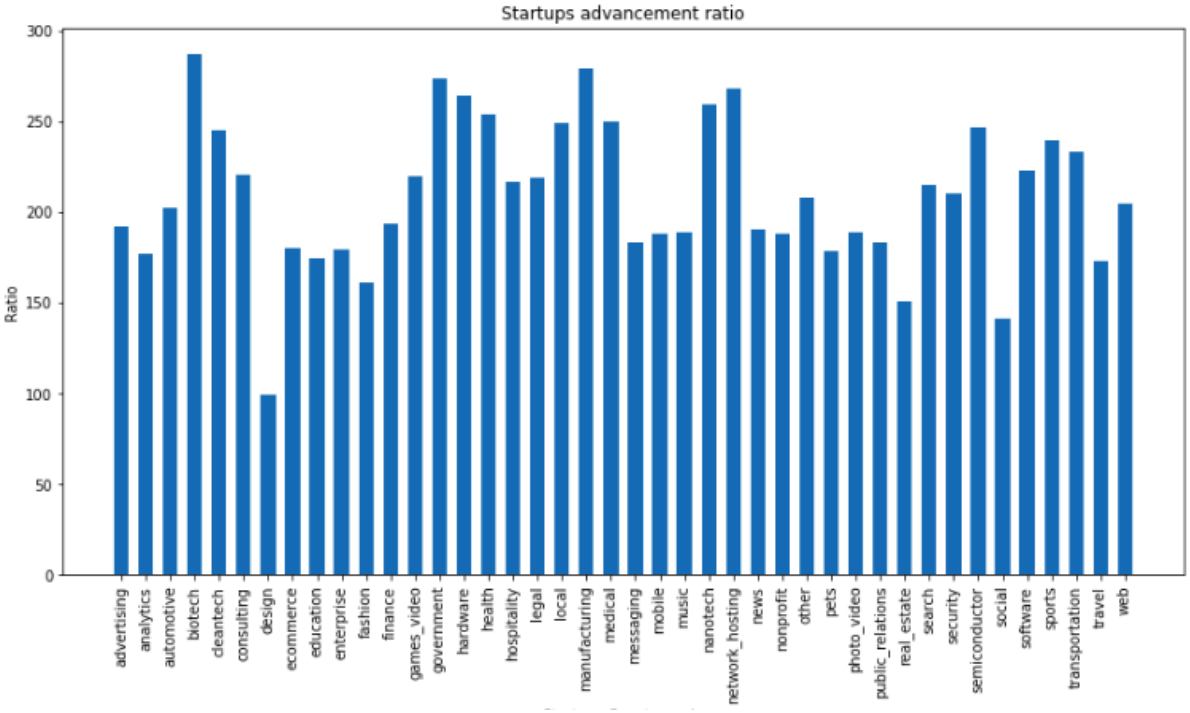
Research Question 2: What is the number of startups in every country?

As shown in the graph below, the distribution of startups in countries, and the USA has the highest number of startups among the countries.



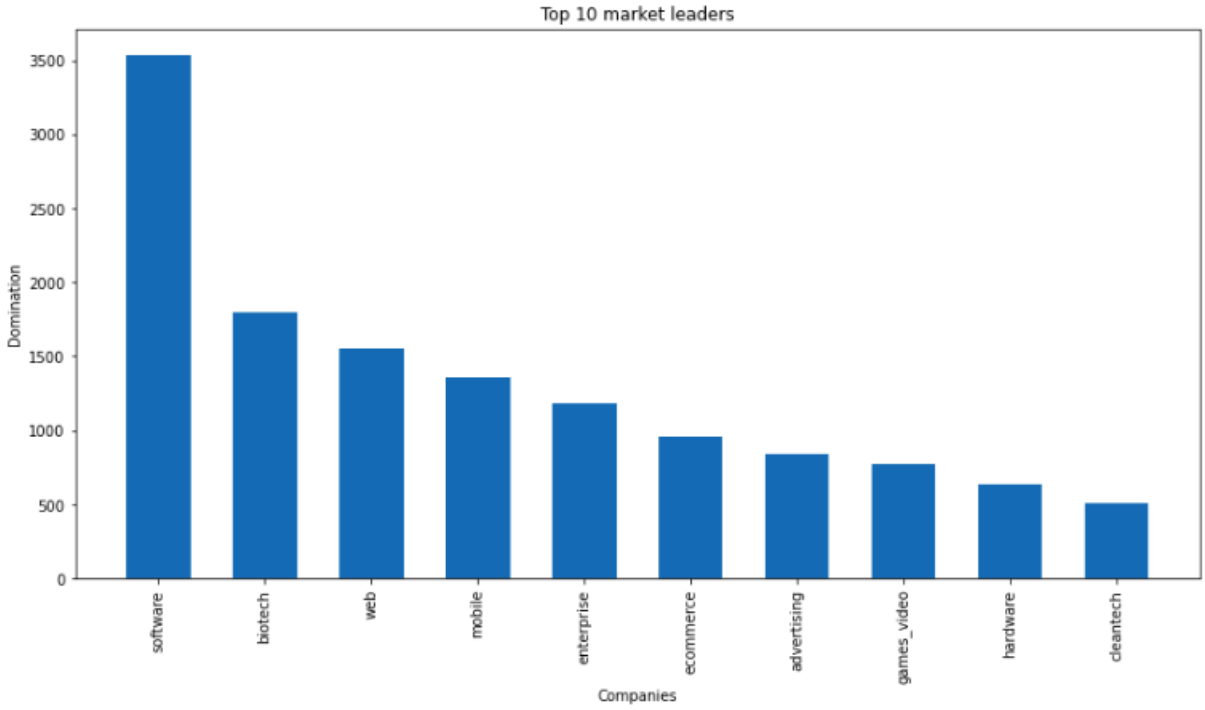
Research Question 3: Which startup's category advances faster?

Biotech is the startup category which advances faster.

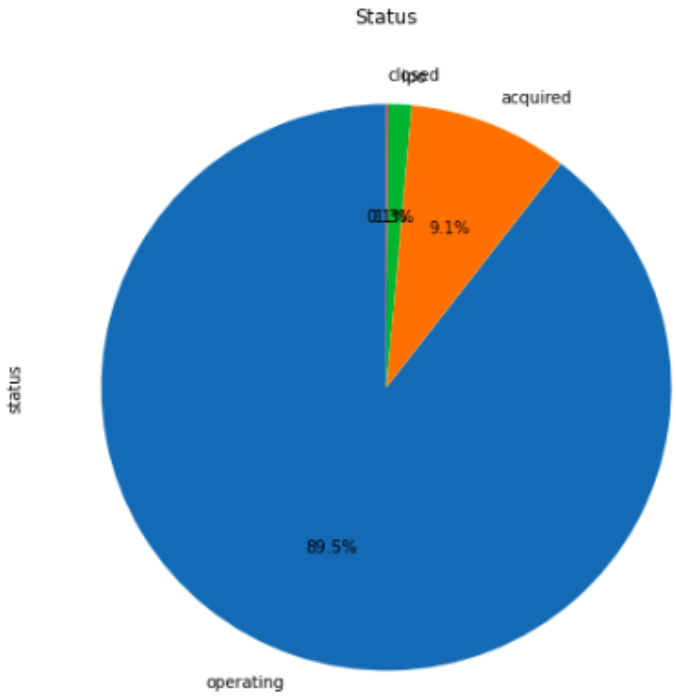


Research Question 4: Who are the top 10 market leaders?

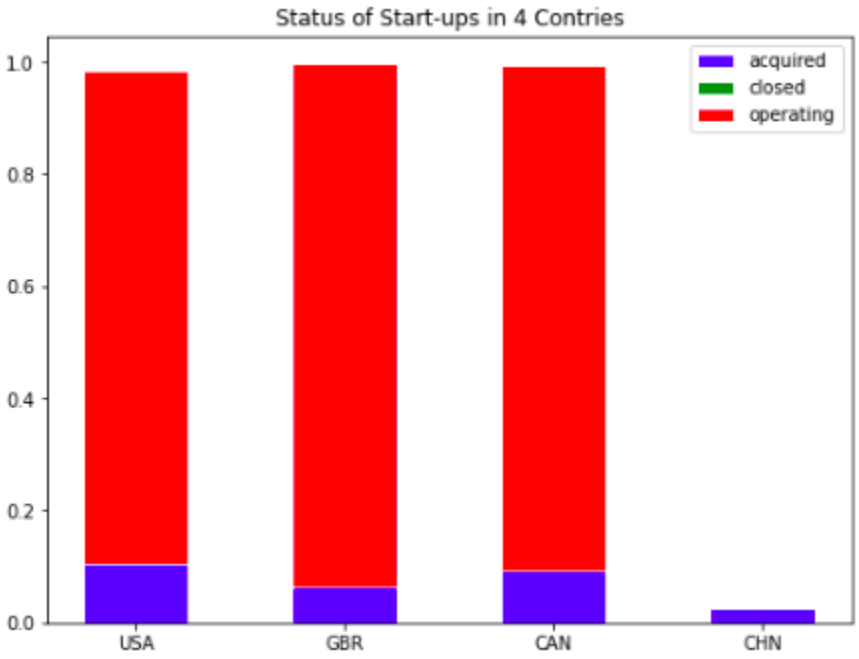
Software, Biotech, Web, Mobile, Enterprise, E-commerce, Advertising, Games videos, Hardware, Cleantech are the top 10 market leaders.



Companies Status



Status of startups in **USA, UK, Canada, China**



Experiment setup

Problems with the dataset

The dataset has imbalanced classes, a problem faced when trying to create a good predictive model for the task at hand was the large class imbalance between operating, ipo, closed, acquired.
After preprocessing, around 3k rows remaining from the dataset consisted of different target classes(operating, ipo, closed, acquired),
Most machine learning algorithms work best when the number of observations of each class is equal because when there is such disparity between classes the algorithms tend to classify the lowest represented class as the opposite.
In the

Machine learning algorithms

In the present work, we have a multi-label classification problem the target is either classified as “1” for operating, “2” for ipo, “3” for “closed”, and “4” for acquired.
It’s a type of supervised learning, a method of machine learning where the output categories are predefined. It is important to choose the algorithm that better fits for the problem but also adapts well to the characteristics of the dataset.
Also we will apply ensemble techniques between the models using VotingClassifier
Different learning algorithms make different assumptions about the dataset and have a different purpose.
When testing the following algorithms we intend to test its data with ML models that not only fit the nature of dataset but are also easy to understand and implement.
In the following sections, we will do an overview of each tested machine learning algorithm- **RandomForest, XGboost**, and will do Ensembling between the two models using **VotingClassifier**.

Random Forest

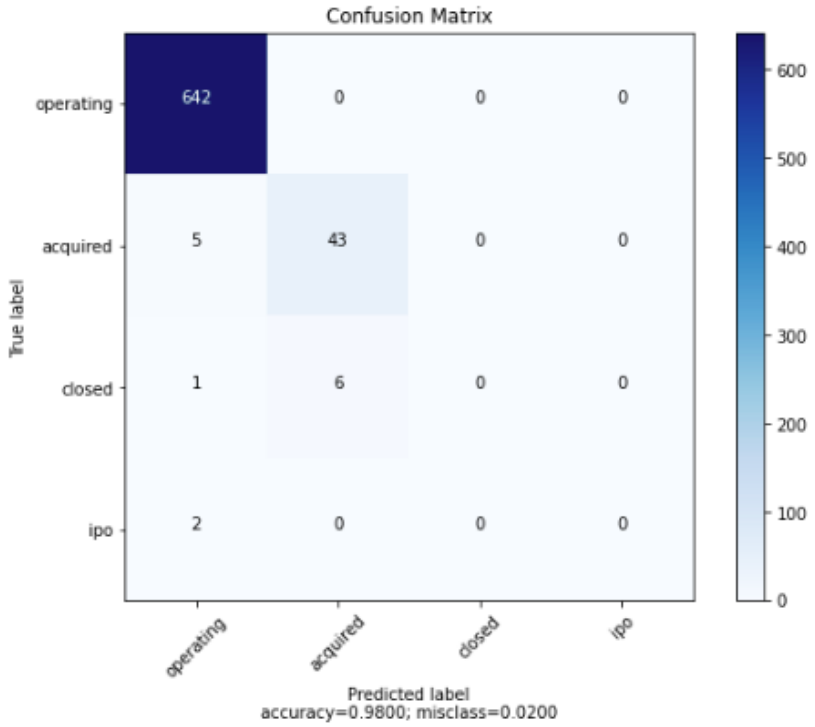
(RF) is a collection of decision trees that can be thought of using bagging on multiple tree classifiers. However, since it is not possible to build multiple trees on the same data as it will get the same results, randomness of two types is introduced: each tree is built on slightly different rows, sampled with repetitions from the original (bagging), and each tree (or in some cases each branch decision) is built using a randomly selected subset of columns. The point of RF is to prevent overfitting, which it does by creating the random subsets of features and building smaller (shallow) trees using the subsets.

The algorithm can be summed up in the following steps.
(1) Draw n tree bootstrap samples from the original data.
(2) For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m of the predictors and choose the best split from among those variables. This sample of m predictors minimizes the correlation between the classifiers in the ensemble.
(3) Predict new data by aggregating the predictions of the n trees (i.e., majority votes for classification, average for regression)

The way the algorithm handles the Bias-variance trade-off, a central problem in supervised learning, is one of the main advantages of Random Forests - although its bias is the same of a single Decision Tree, its variance decreases as we increase the number of trees which also decreases the chances of overfitting. Other advantages are the fact that it runs efficiently on large datasets, handles thousands of input features without feature deletion, gives estimates of what variables are important to the classification, processes missing data and even maintains high accuracy when this proportion is large.

The main disadvantage of Random Forests compared with a simple Decision Tree is its interpretability as it is hard to see the relation between a dependent variable and the rule set created. A Random Forest must be a predictive tool and a descriptive one. It is easy to see its features importance but that might not be enough when the objective of the study is to understand the relationship between dependent and independent variables.

The model confusion matrix on our cleaned dataset:



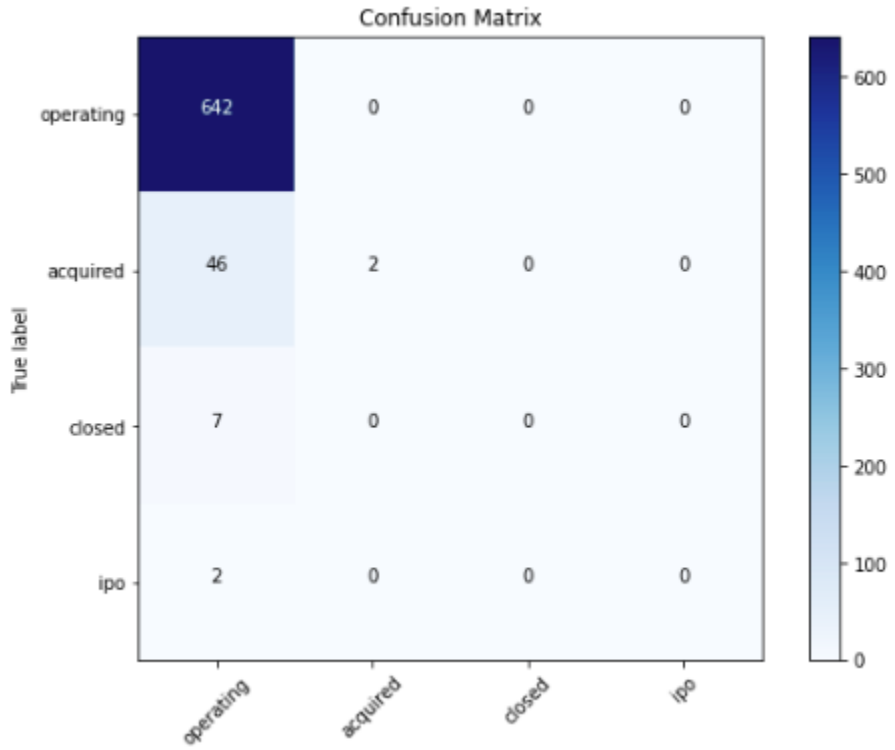
XGBoost(e**X**treme **G**radient **B**oosting)

XGBoost is an algorithm that has recently been dominating applied machine learning and kaggle competitions for structured or tabular data.
XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.
The implementation of the algorithm was engineered for efficiency of compute time and memory resources. A design goal was to make the best use of available resources to train the model. Some key algorithm implementation features include:

- Sparse aware implementation with automatic handling of missing data values
- Block structure to support the parallelization of tree construction
- Continued training so that you can further boost an already fitted model on a new data

The two reasons for using XGBoost are that it is fast when compared to other implementations of gradient boosting.

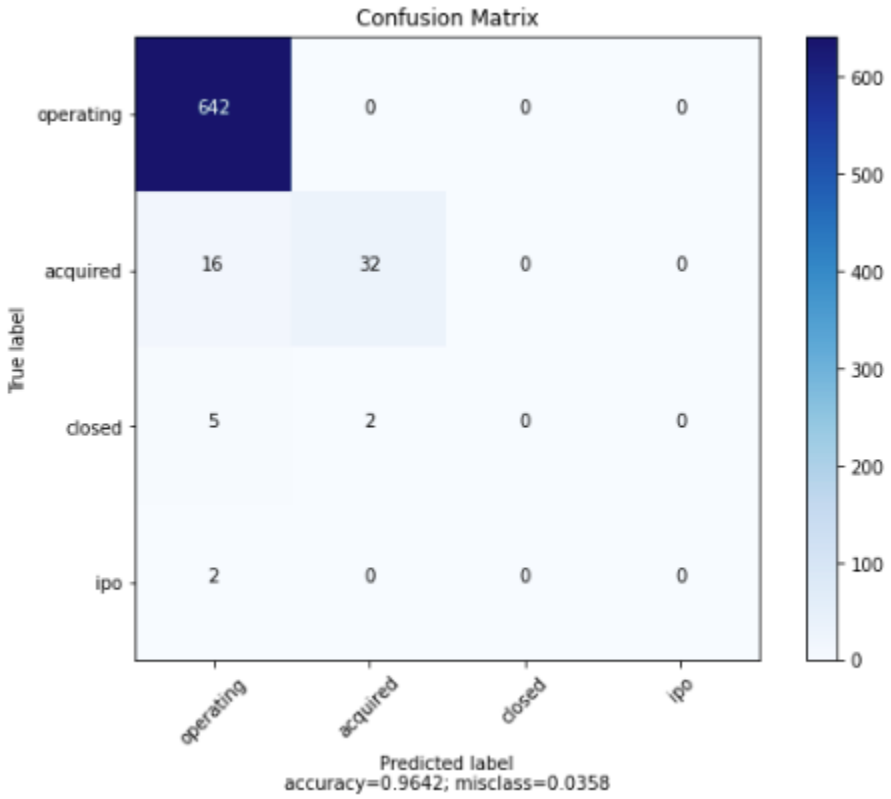
The model confusion matrix on our cleaned dataset:



VotingClassifier

A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.
It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each of them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

We performed a Voting classifier between our models to get the best results we can use for predictions. And its confusion matrix was as follows:



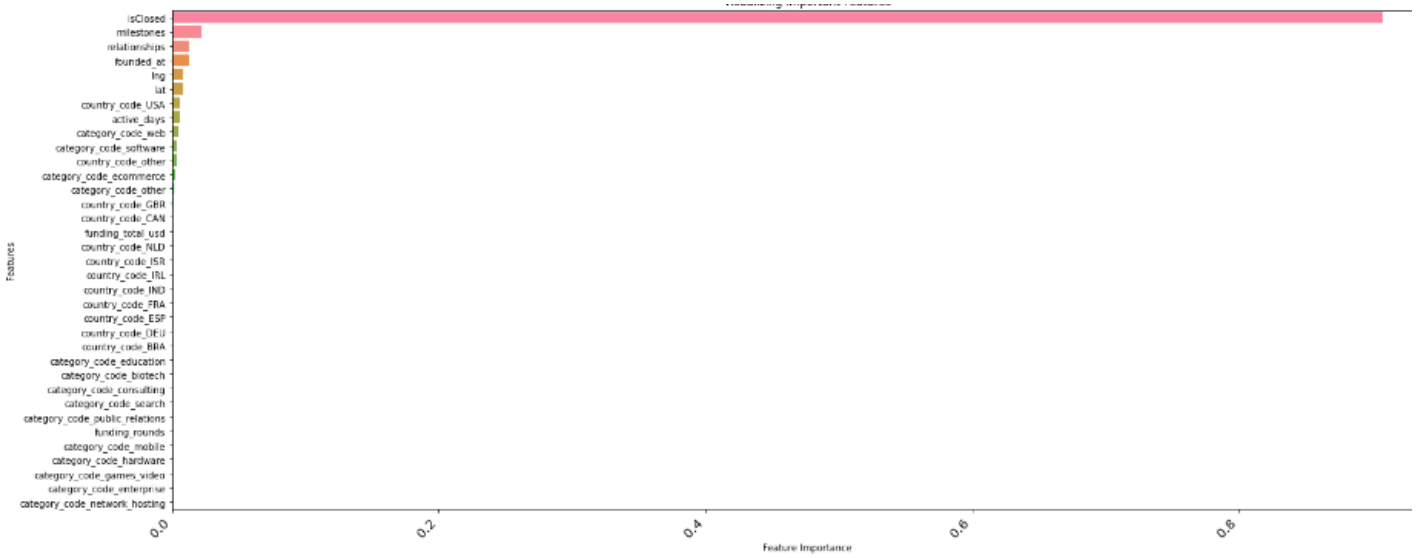
Feature Importance

Machine learning (ML) algorithms such as neural networks and Random Forests (RF) or any other machine learning algorithms are often considered to produce black box models because they do not provide any direct explanation for their predictions. However, these methods often outperform simple linear models or decision trees in predictive performance as they can model complex relationships in the data.
Feature importance used to clarify how the features affects the model performance, to know what is the most effective features to our model.
The Table below shows the results of the selected features from the data cleaning part and used to train the model of VotingClassifier.

Features	**Importance**
isClosed	0.907968
milestones	0.022310
relationships	0.012279
founded_at	0.012207
lng	0.008590
lat	0.008111
country_code_USA	0.005945
active_days	0.005616
category_code_web	0.005260
category_code_software	0.003442
country_code_other	0.003367
category_code_ecommerce	0.002235
category_code_other	0.001708
country_code_GBR	0.000960
country_code_CAN	0.000000
funding_total_usd	0.000000
country_code_NLD	0.000000
country_code_ISR	0.000000
country_code_IRL	0.000000
country_code_IND	0.000000
country_code_FRA	0.000000
country_code_ESP	0.000000
country_code_DEU	0.000000
country_code_BRA	0.000000
category_code_education	0.000000
category_code_biotech	0.000000
category_code_consulting	0.000000

category_code_search	0.000000
category_code_public_relations	0.000000
funding_rounds	0.000000
category_code_mobile	0.000000
category_code_hardware	0.000000
category_code_games_video	0.000000
category_code_enterprise	0.000000
category_code_network_hosting	0.000000

Visualization of the effect of the features on the model performance:



Deployment

Heroku is a cloud platform as a service (PaaS) supporting several programming languages. One of the first cloud platforms. Also it lets companies build, deliver, monitor and scale apps in a fast way to go from idea to URL, by passing all those infrastructure headaches.

How our application looks like using simple HTML and CSS.
It consists of 2 forms, one for submitting startups' information and the other one for showing the results.

Predicting Startup's Acqusition Status

Enter startup information

Year of foundation

Closed at (year)

if not closed skip it

Location

Funding rounds

Total funding in USD

Milstones achieved

Number of relations

Category Code

Country Code

Submit

Results

Startup is Acquired

Conclusion

The main objective of the present study was to generate a model to classify companies or Start-ups whether they are operating, ipo, closed, or acquired. By building multilabel classifiers to classify startups with a AUC of 96.42% it is assumed that the objective was achieved. It is the highest reported using data from CrunchBase. The model can classify with high efficiency not only the total of successful companies in the dataset (TPR, recall) but also, from all the successfully-classified which are successful (Precision). The machine learning algorithm used is VotingClassifier which provides a fast and easy to interpret and implement model with positive results. It provided better results than running XGBoost or RandomForest separately. its potential to fit in the size and nature of our dataset as a linear relation was expected.
During the experiment setup, a transformation on all features was tested. (at the cost of higher time to compute). This transformation allows the model to pick features which are specific values of the feature – allowing the model to learn from more specific information (through a higher number of combinations between features).

Future Work

- Fix the problem of imbalanced data
- Add more HTML and CSS for the project
- Optimize models accuracy

References

https://medium.com/m/global-identity?redirectUrl=https%3A%2F%2Ftowardsdatascience.com%2Fexploratory-data-analysis-8fc1cb20fd15
https://en.wikipedia.org/wiki/Exploratory_data_analysis
https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/
https://en.wikipedia.org/wiki/Heroku
Fortune 1000 Companies List for 2016 - Geolounge. (n.d.). Retrieved May 23, 2017, from https://www.geolounge.com/fortune-1000-companies-list-2016