

CIND820: Big Data Analytics Project

Iuliia Sharafutdinova

D1H & 501188055

Uzair Ahmad, Ph.D

November 21, 2022

Table of Contents

Abstract.....	3
Literature Review.....	4
Data Description.....	6
Approach.....	16
Initial Results and Code.....	18
References.....	25

Abstract

Over the past decade, billions of dollars have been invested by funds, government and financial institutions in entrepreneurial ventures—what is often referred to as venture capital. This project aims to analyze attributes of venture capital investments in start-up industry by using a panel data set of 101 countries over the period 1983 to 2014 in order to build predictive model for success and failure raising funds.

This study analyzes critical factors of success and failure of startups. The main goal, therefore, is to identify which crucial characteristics found in the sample had the greatest impact on the success and failure of the startup to receive funding.

The dataset about start-up companies is extracted from CrunchBase dated December 02, 2014 and downloaded from Data.World website: <https://data.world/datanerd/startup-venture-funding>.

The dataset includes amount invested by funds, acquisitions, stages, the total amount of venture capital, funding rounds, regions, categories of the companies and year of foundation. The dataset is embracing 114,506 number of rows and 24 attributes including both 1 quantitative and 23 qualitative target attributes.

Four machine learning methods will be used in this project – Decision Tree, Random-Forest and KNN.

The Decision Tree and Random-Forest will be created using Python and R, while the KNN method will be applied through Python. Finally, Tableau will be used for designing visualization of analysis results.

Literature Review

The global venture capital market has been grown up significantly. Notably millions of startups in the world have been backed by venture capital. For example, according to the KPMG International Global analysis of venture funding report 2022, amount of venture capital investment grew from about US\$ 337.3 billion in 2018 to US\$ 671 billion invested in 2021, growing at the average annual rate of 50% over the past three years. The robust venture capital investment climate was highlighted by record setting investment levels in numerous jurisdictions, including the US, Canada, Brazil, the UK, Germany, Israel, the Nordic region, Ireland and India.

Lerner and Tag (2013) emphasized that the ability of venture capital investors to overcome information asymmetries and provide capital to the most innovative and successful startups boosted by economic growth only. Allen (2012) mentioned the positive effect of welfare on venture capital investment. In addition, Popov and Roosenboom (2014) found that venture capital increases the rate, the quality and quantity of new business creation as well as innovative products and services. Although positive effects of venture capital investments to society is described extensively in academic research, however, critical factors that led to success or failure in venture capital attraction by one startup firm or another have not investigated yet.

Paul Gompers and Josh Lerner (2001) addressed such factor - the understanding of the Internet as tool and technology by big corporations. That factor led to rapid growth of venture fundraising in 1990s by expanding the Internet and an understanding of its implications which eventually, triggered the increasing corporate interest and investments in venture capital. Although a wide range of media, service, and manufacturing firms realized the potential of the Internet to challenge their traditional ways of doing business—but they had few internal resources to address the new communications technology.

KPMG Global analysis of venture funding report (2022) summarized a few additional factors. According to this report the recent explosive growth of the global venture capital market was conditioned the ready availability of cash, the significant returns seen on exits throughout the year, and the increasing participation of corporates, family offices, and a range of other non-traditional investors has only added to the overall attractiveness of the venture market. The combination of a

strong investment environment and the continued drive for digitalization keep venture investments high heading.

William Janeway, Ramana Nanda and Matthew Rhodes-Kropf (2021) documented that governments play the important role of raising and stimulating venture capital too: the billions of dollars that have been committed to help support venture-backed firms in Europe, North America and beyond since the onset of the COVID-19 crisis underscore the policy interest in venture capital. This is another important source and factor of explosive growth in venture capital investments. Therefore, startups raised funds not only because of business or cash need but also because of availability of venture capital.

Bruno and Cooper (1982) highlighted that regardless of economic growth, government assistance and other institutions, the majority of new companies cannot raise money from venture capital investors as result go bankrupt. According to their study of 250 Silicon Valley new technology firms founded in the 1960s, 36.8% had been discontinued, 32.4% had merged or had been acquired, and only 30.8% had survived as independent companies. Moreover, Juan B. Rour and Modesto A. Maidique (1986), pointed out factors that led to success: founders' track record based on previous education, work experience, team management skills as well as targeted market for a new product or a service. They also mentioned that market share targeted by a new company will be positively correlated with success of the capital venture to attract capital.

Data Description

The dataset consisted of 114506 rows and 24 attributes – 1 of which were quantitative and 23 which were qualitative including the target attribute. Table 1 shows the attributes and the attribute type.

Table 1. Attribute Type and Description

Column Name	Attribute type	Description
company_permalink	nominal	type of persistent identifier the company
company_name	nominal	name of company
company_category_list	nominal	categories describe the business
company_market	nominal	categories describe the market of business
company_country_code	nominal	name of country for company
company_state_code	nominal	name of state for company
company_region	nominal	name of region for company
company_city	nominal	name of city for company
investor_permalink	nominal	type of persistent identifier the investor
investor_name	nominal	name of investor
investor_category_list	nominal	categories describe the investor
investor_market	nominal	categories describe the market of investor
investor_country_code	nominal	name of country for investors
investor_state_code	nominal	name of state for investors
investor_region	nominal	name of region for investors
investor_city	nominal	name of city for investors
funding_round_permalink	nominal	type of persistent identifier for rounds
funding_round_type	nominal	type of round
funding_round_code	ordinal	code of round
funded_at	ordinal	funded data
funded_month	ordinal	funded month
funded_quarter	ordinal	funded quarter
funded_year	ordinal	funded year
raised_amount_usd	quantitative	raised amount in USD

Table 2 shows the min, max, mean, median, and standard deviation of the numeric attributes. If the mean is closer to the minimum value, then the data is right-skewed, if it is closer to the maximum value then it is left-skewed, but if it is closer to the median then it is normally distributed. Based on the values, the attributes raised_amount_usd have a mean that is close to the maximum so it should be left-skewed.

Table 2. Numerical Attributes Data Exploration

Attribute	Min	Max	Mean	Median	Standard Deviation
raised_amount_usd	0	5.800e+09	1.119e+07	3.064e+06	4.724041e+07

The dataset does not have any duplicate records, so no rows were deleted for this reason. The dataset has duplicate columns that do not significant contribution or influents on results. These columns are excluded: funded_at, funded_month, funded_quarter, funding_round_permalink.

The data was checked for missing values. From data set company_state_code, investor_category_list, investor_market, investor_state_code and funding_round_code have been removed as these columns have more than 30% null values. Table 3 shows number of missing values in dataset.

Table 3. Missing values

Column Name	Amount	Proportion (%)
company_permalink	0	0.00
company_name	1	0.00
company_category_list	3264	2.85
company_market	3266	2.85
company_country_code	7359	6.43
company_state_code	35348	30.87
company_region	7359	6.43
company_city	8705	7.60
investor_permalink	66	0.06
investor_name	66	0.06
investor_category_list	83999	73.36
investor_market	84051	73.40
investor_country_code	27958	24.44
investor_state_code	52232	45.63
investor_region	27958	24.44
investor_city	28499	24.89
funding_round_permalink	0	0.00
funding_round_type	0	0.00
funding_round_code	59837	52.26
funded_at	0	0.00
funded_month	0	0.00
funded_quarter	0	0.00
funded_year	0	0.00
raised_amount_usd	13351	11.66

Outliers are values that are outside the normal range or at an abnormal distance from other points in a dataset (Frost, n.d.). It is important to identify and deal with outliers because they can be errors or an anomaly which can distort the analysis of the data (Barnes, 2021). Identifying and dealing with them is a way of confirming the data quality before performing any further analysis or

extracting insights (Barnes, 2021). Some of the ways to determine outliers are using graphs such as box plots, histogram or a scatter plot, and z-score (Frost, n.d.).

Outliers can be dealt with by removing the records with the outliers or transforming the data using log or square root. Removing records with outliers may not be an ideal solution in every case because of the loss of data that comes with it. Also, outliers should be investigated before deleting them from the dataset because they could be valid (Grace-Martin, n.d.). An alternative is to remove the attributes with a significant proportion of outliers but first, you must determine if the attribute is a key feature.

The box plot (univariate method) was used to determine the outliers for the quantitative attributes in this dataset. The multivariate method can also be used to determine the outliers. This method might be more accurate since it builds a model using all the attributes (Alberto Quesada, 2017).

Table 4 below shows the outliers for the quantitative attributes. From the table, `fraised_amount_usd` has 8.4% outliers. Since data manipulation is out of the scope of this project and to prevent loss of data, no records with outliers were remove.

Table 4. Numerical Attribute Outlier Analysis

Attribute	# of Outliers	% Outlies
<code>fraised_amount_usd</code>	9627	8.4

Correlation indicates the relationship between 2 or more variables. Table 5 shows the correlation between numerical attribute `raised_amount_usd` and ordinal `funded_year`. The table illustrates how much funds raised per year in million dollars. According to the table, year 2014 had the largest portion of venture capital investments among all countries. In monetary terms, in 2014 investments reached \$5.8 billion, which was more than twice than in 2013 (\$2.6 billion). Notably, regardless of financial crisis, the amount of venture capital investments peaked to \$3.2 billion in year 2008. Overall, there was upward trend in venture capital investment worldwide since 1990s until 2014.

Table 5. Funds raised per year in million dollars

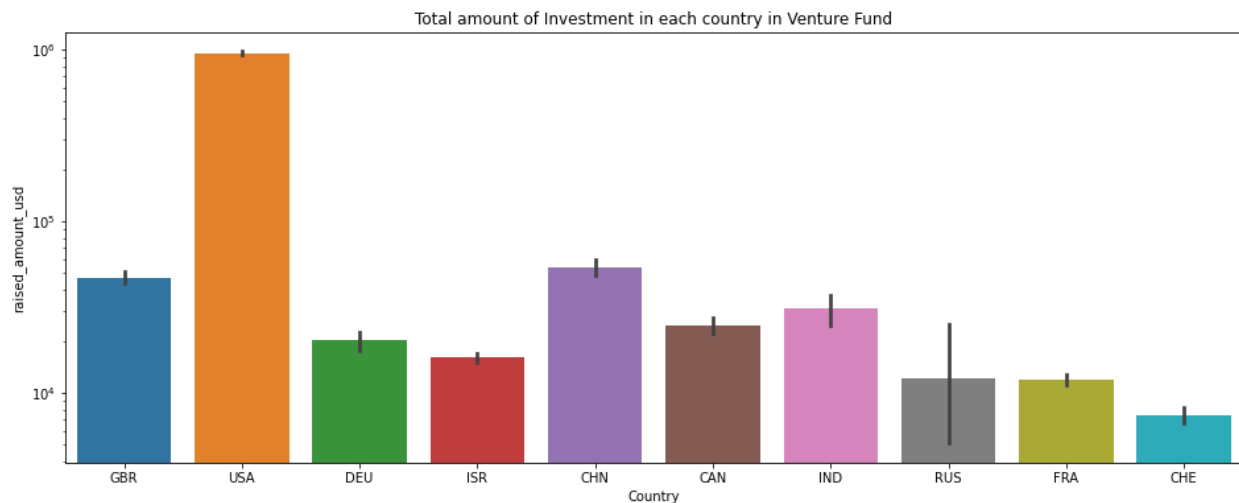
funded_year	mean	median	min	max
1921	0.001000	0.001000	0.001	0.001
1974	0.000000	0.000000	0.000	0.000
1979	1.000000	1.000000	1.000	1.000
1982	0.348000	0.165000	0.155	0.724
1983	0.094000	0.094000	0.094	0.094
1984	0.000000	0.000000	0.000	0.000
1985	0.215600	0.169000	0.000	0.666
1986	0.000000	0.000000	0.000	0.000
1987	0.425143	0.000000	0.000	2.500
1988	0.000000	0.000000	0.000	0.000
1989	0.000000	0.000000	0.000	0.000
1990	6.468947	6.468947	0.000	17.550
1991	0.000000	0.000000	0.000	0.000
1992	0.226750	0.000000	0.000	2.000
1993	0.000000	0.000000	0.000	0.000
1994	2.755625	0.000000	0.000	13.000
1995	1.552941	0.000000	0.000	8.000
1996	2.434333	0.000000	0.000	42.000
1997	5.103750	1.850000	0.000	50.000
1998	2.998523	0.000000	0.000	25.000
1999	8.363284	3.000000	0.000	91.000
2000	19.065850	10.500000	0.000	120.000
2001	9.902324	4.000000	0.000	90.000
2002	8.956812	5.000000	0.000	200.000
2003	7.052222	4.600000	0.000	75.000
2004	10.085691	6.000000	0.000	110.000
2005	11.915911	8.000000	0.000	1000.000
2006	11.683749	7.500000	0.000	422.000
2007	12.688934	7.000000	0.000	400.000
2008	15.463420	6.500000	0.000	3200.000
2009	12.482388	5.000000	0.000	1500.000
2010	9.688145	3.230000	0.000	565.000
2011	11.383164	2.000000	0.000	1500.000
2012	7.963645	1.602564	0.000	1050.000
2013	8.563118	1.700000	0.000	2600.000
2014	14.504937	2.723336	0.000	5800.000

Samuel Kaski (1997) highlighted that most exploratory data analysis techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of data analysis is to explore. One part of exploratory data analysis is visualized potential relationships (direction and magnitude) between variables.

Bar chart 1 illustrates total amount of venture capital investments across top largest 10 countries. According to the bar chart, the USA start-ups received the largest portion of venture capital invested among all countries between 1974 and 2014. In monetary terms, the USA accumulated about \$ 947,912 billion, however, only \$54,054 billion invested in venture capital industry in China, followed by Great Britain with \$47,044 billion. Therefore, United States venture investments alone is representing more than half of overall amount invested in the entire industry for 40 years.

Table 6. The total investment across top 10 countries

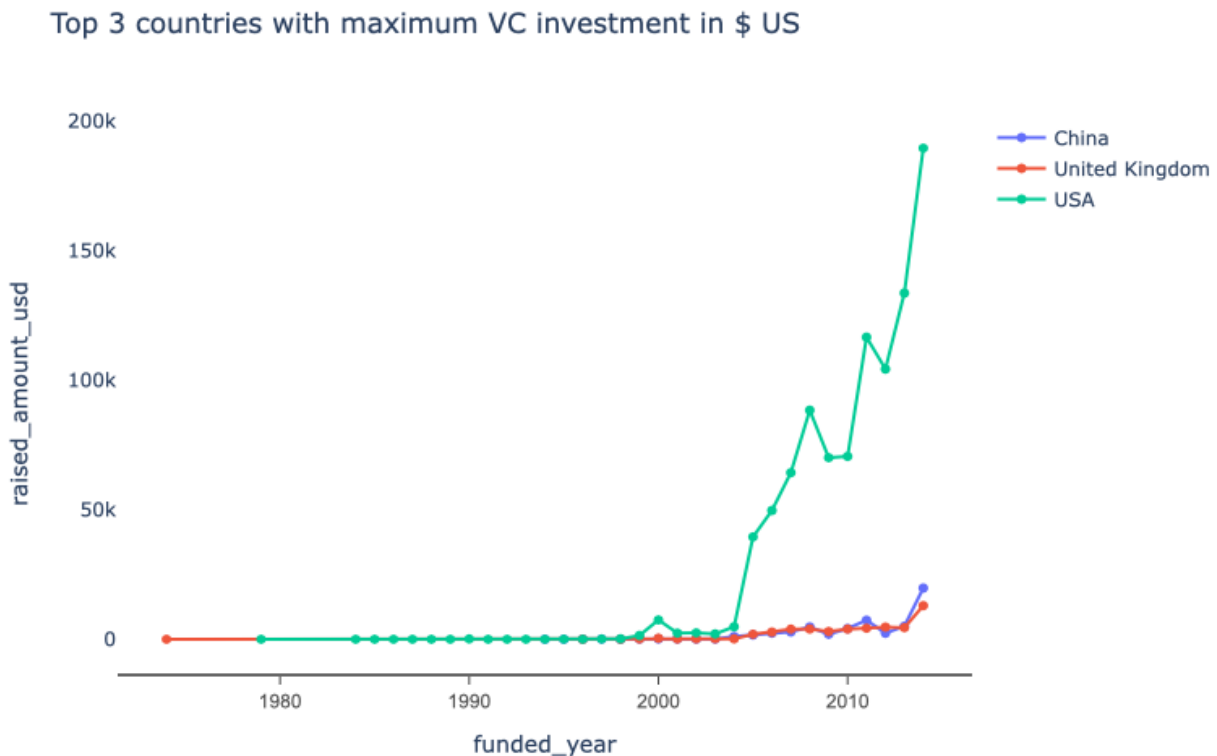
company_country_code	raised_amount_usd (in million)
United States of America	947 912.862387
China	54 054.587503
The United Kingdom	47 044.416507
India	30 824.652715
Canada	24 666.404858
Germany	20 154.017753
Israel	16 010.645140
Russia	12 114.162814
France	11 982.181676
Switzerland	7 383.477633



Bar Chart 1. Total amount of Venture Capital Investment in each country (1974 - 2014)

Line chart 1 shows the dynamic of investments across the top three countries USA (green), China (blue) and UK (red) since year 1980.

It is seen on this line chart, there is significant growth of venture capital investments in the US in 2004, which peaked to \$88 billion in 2008. Then, the amount of funds raised by startup firms had declined to \$70 billion due the global financial crisis in 2008. Nevertheless, venture capital industry recovered after 2010 and number of investments went up to \$116 billion. Notably not only the funding in the U.S. business startups had declined in 2012 but also, venture capitalists spent less money on fewer deals. Finally, US investments grew from \$104.30 billion to \$189.57 billion between 2012 and 2014. In contrast, UK and China were lagging behind in both monetary and quantity of venture capital deals since 1999. It is concluded that the United States remains the largest recipient of venture capital investments.

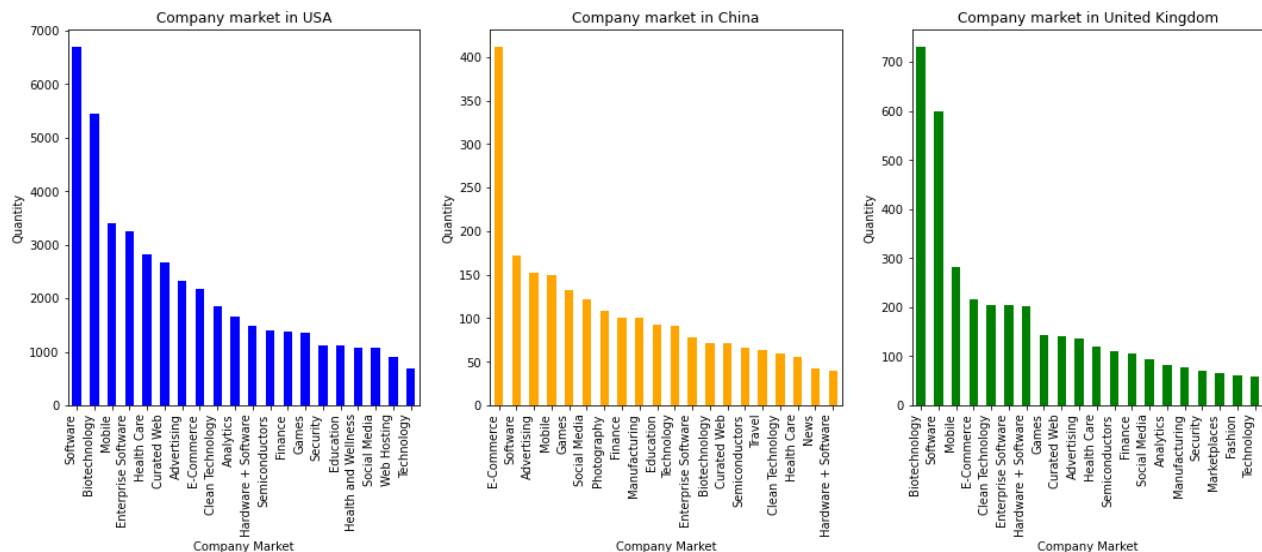


Line chart 1. Top 3 countries with maximum investment in USD

Next it is important to analyze venture financing in terms of sectors. The bar chart below presents the 20 economic sectors with the largest amount of venture capital investment. In the US, Software sector is the largest recipient of funding, which equals to 6687 investments. The second sectors are Biotechnology and Mobile 5444 and 3392 respectively.

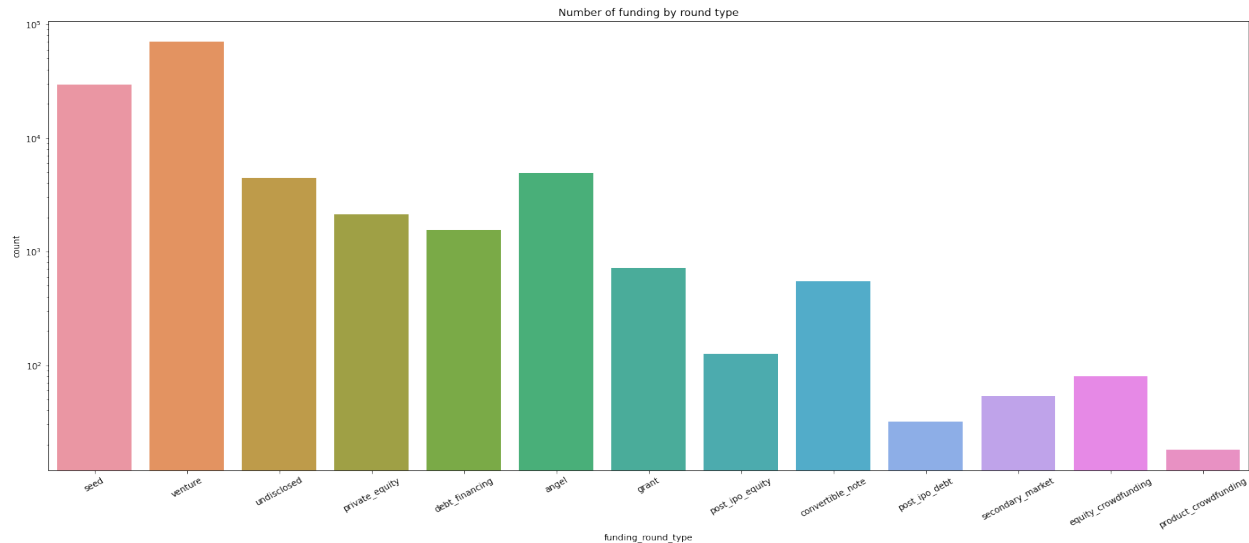
As for China, venture capital investors had targeted E-Commerce (411), Software (172) and Advertising (152) services sectors.

Biotechnological startups had received the most of funds reaching 730 investments, followed by Software with 600 and Mobile with 282 deals. However, firms from other economic sectors include information technology services, health care, game and clean technology had not attracted much of venture capitalists' attention.



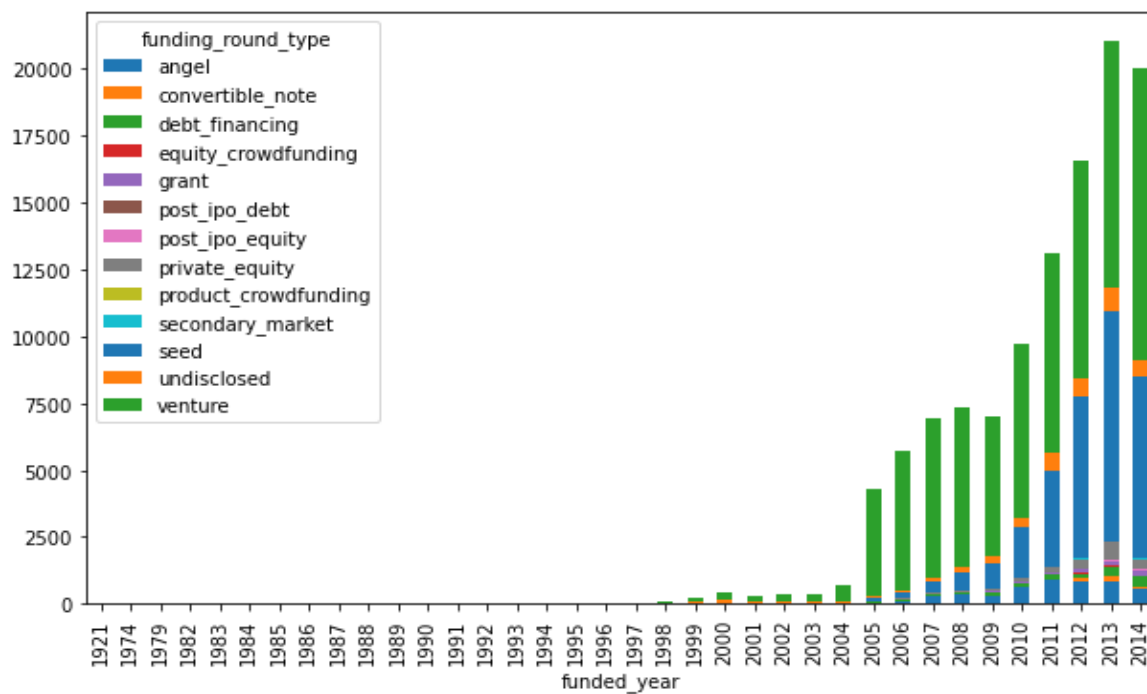
Bar chart 2. Group by company market to see which market has highest number of investments in US, UK and China

Bar chart 3 presents the numbers of startups by each round: venture, seed, undisclosed, convertible_note, private_equity, debt_financing, angel, grant, equity_crowdfunding, post_ipo_equity, post_ipo_debt, product_crowdfunding, secondary_market, non_equity_assistance versus number of investments. It indicates that post_ipo_equity and product_crowdfunding are the least successful to attract investment. However, venture, seed and angel rounds during the earliest stages of startup lifecycle has been the biggest in terms of amount invested. This can be partially explained by lower barrier and amount threshold to invest in startups at these stages.



Bar chart 3. Number of funding by round type

Stack chart 1 visualizes the numbers of startups for each round between 1998 and 2014. The average amount by funding round has increased significantly in 2006 years in the whole world. The largest part is represented by both venture round and seed types. Notably, the biggest investment amount was in 2013 year.

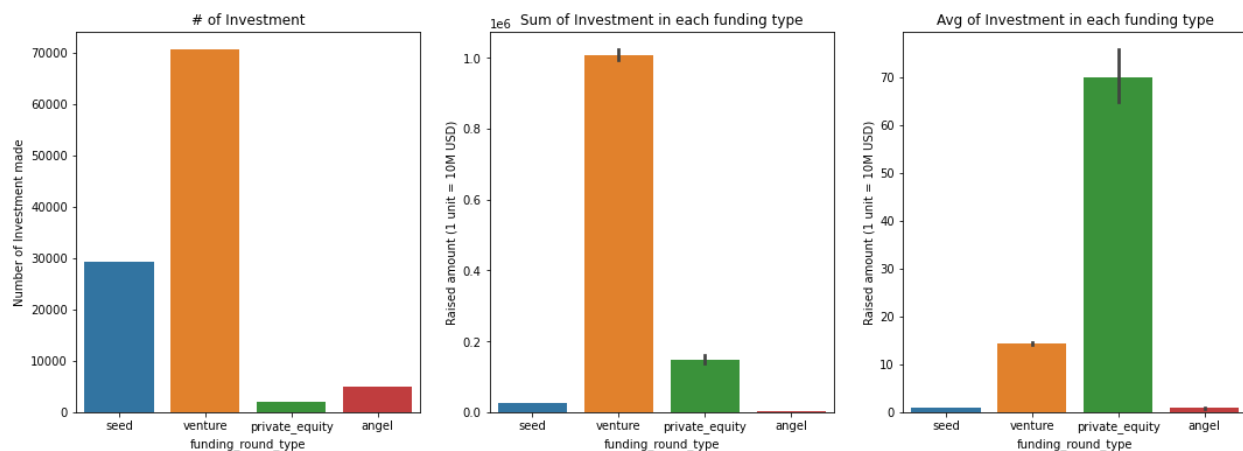


Stack Chart 1. Distribution of funding by round type (1921 - 2014)

Bar charts 4 illustrates four funding types: Venture, Seed, Private Equity and Angel. First left chart visualizes the numbers of investments for each round, which are venture round 70,615, seed 29,272, angel 4 894 and private_equity 2,128.

Middle chart shows the sum of investment in each funding type. It is seen from this chart that the highest round is venture about \$10 billion and the smallest are private equity, angle and seed. Right chart shows the average of investment in each funding type. The biggest amount is the private_equity stage, about \$69.85 million, followed by venture funding stage is average \$14.27 million, finally seed and angel funding average is approximately \$0.9 million each.

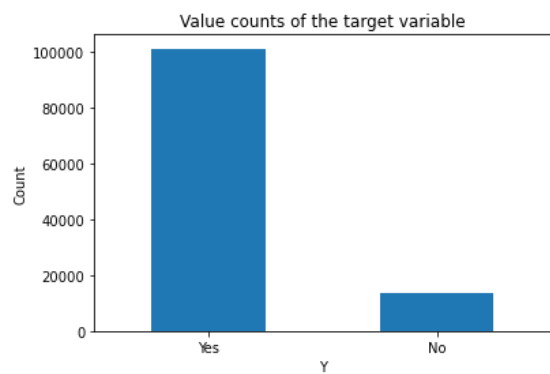
Such difference can be explained by fact that seed and angel funding refer to early-stage startups whereas venture funding occurs after seed or angel stages and involves a relatively higher amount of investment. Private equity type investments are associated with larger companies and involve investments bigger than venture type. Startups which have grown in scale may also receive private equity funding.



Bar charts 4. Analysis of selective funding round type - Venture, Seed, Private Equity and Angel

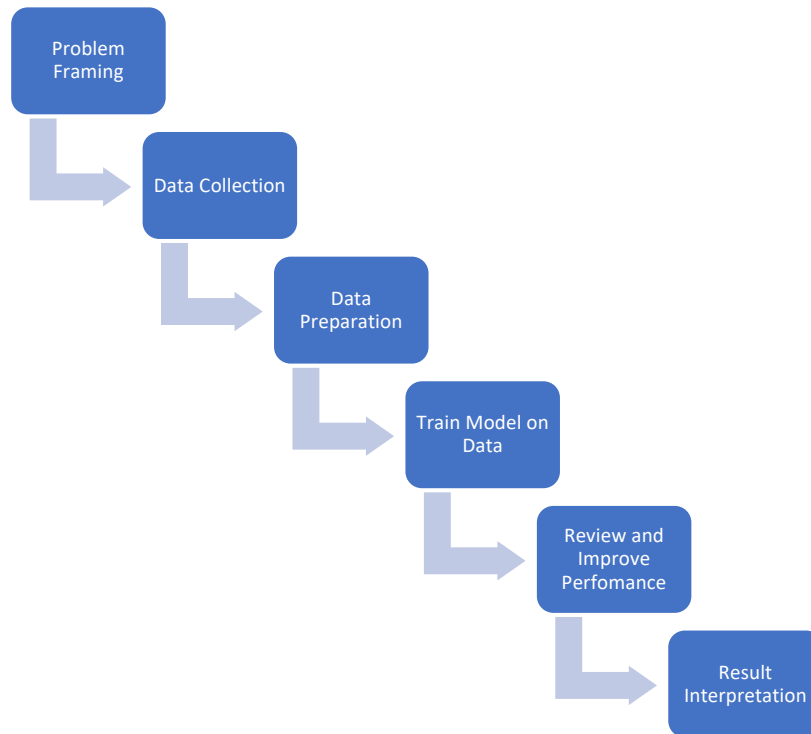
The biggest determinant of appropriate algorithms and methods to use with a dataset is the target variable. The target variable, also known as the dependent variable, is what trying to predict.

Target variables is Y and it has been added and converted from numeric raised_amount_usd to categorical Y. Graph 5 shows distribution of Y, it is right-skewed. Variable “Yes” near 88.2 % and “No” is 11.7%.



Bar charts 5. Value counts of the target variable

Approach



Schema 1. Approach of the project

Defining problem is the process of analyzing a problem to isolate the individual elements that need to be addressed. Problem framing helps to determine project's technical feasibility and provides a clear set of goals and success criteria. When considering an ML solution, effective problem framing can determine whether your product ultimately succeeds. The main goal is to identify which crucial characteristics found in the sample that has the greatest impact on the success and failure of the startup to receive funding.

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables answering stated research questions, test hypotheses and evaluate outcomes.

The data has been acquired from a public source Data.World website: <https://data.world/datanerd/startup-venture-funding>. The data contains information about startup

companies, investments, and acquisitions via Crunchbase. csv.reader is uploaded to enable manipulation of data in Python.

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. This process purpose is to identify bigger and the most important trends and major points to conduct analysis. Exploratory data analysis has been applied for data to learn insights about investment data further. Pandas, numpy, matplotlib, plot, seaborn libraries have been used to perform operations with rows or columns and visualize data. These steps are required to identify features and label sources.

The next step is building a machine learning model. This process identifies as training model over a set of data, providing an algorithm to analyze and learn from the dataset. Furthermore, the type of model is identified. The difference depends on the type of task the model needs to perform and the features of the dataset at hand.

The model to perform ML are Supervised Learning, Classification. Classification is a supervised machine learning technique used to train models which predict classes based on the training from the training dataset. In this project, the Decision Tree and Naïve Bayes classification techniques are used to train, test and evaluate the performance of the dataset. After applying the algorithm, the next step is to find the accuracy of the model.

Review and improve performance stage is the process that helps to analyze and improve results of machine learning model. This process ensures that the quality of the data is accurate at various stages of analysis. Also, it assists to verify that expected results are plausible and fairly presented.

Result interpretation is the process of using diverse analytical methods to review data and arrive to relevant conclusions.

The result and code for the project is uploaded on Github: <https://github.com/sharafjul/CIND-820-Project> .

Initial Results and Code

Mitchell (1997) defined Machine Learning (ML) as a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.

The first step of any machine learning project is developing an understanding of the goal of this project. This goal is to identify which crucial characteristics found in the sample that has the greatest impact on the success and failure of a startup to receive funding from dataset. One of the most important steps is to preprocess data. The preprocessing allows the data to be understandable by the machine learning algorithms. Preprocessing can be split into cleaning, normalizing, transforming, imputation of missing values, integrating, and identification of noise in data.

Then the project next steps are Feature Selection and Data Partitioning. According to Johnson and Kuhn (2013) feature selection is primarily focused on removing non-informative or redundant predictors from the model. Therefore, `company_permalink`, `company_state_code`, `investor_permalink`, `investor_category_list`, `investor_market`, `investor_state_code`, `funding_round_permalink`, `funding_round_code`, `funded_at`, `funded_month`, `funded_quarter` attribute were removed from dataset during the exploratory data analysis because one part of the columns has more than 30% null values and other part is not significantly contributing or influencing results.

Chi-Square test has been applied for feature selection of remaining values. The purpose of this test is to identify whenever the feature is categorical. However, the target variable is defined as categorical. Python package has been used to run Chi-Square test. Following data attributes/parameters were selected `company_name`, `company_category_list`, `company_market`, `company_country_code`, `company_region`, `company_city`, `investor_name`, `investor_country_code`, `investor_region`, `investor_city`, `funding_round_type`, `funded_year`, `raised_amount_usd` to reject the null hypothesis.

These codes and results can be found at GitHub: <https://github.com/sharafjul/CIND-820-Project/blob/main/Feature%20Selection%20via%20Chi-Square.ipynb>

Table 7. Results of Feature Selection via Chi-Square.

Column	Chi2 Statistic:	p-value:	Hypothesis
company_name	58995.4264	0.0	Reject Null Hypothesis
company_category_list	26232.8076	0.0	Reject Null Hypothesis
company_market	2884.5785	0.0	Reject Null Hypothesis
company_country_code	2801.5557	0.0	Reject Null Hypothesis
company_region	5547.6830	0.0	Reject Null Hypothesis
company_city	9563.1697	0.0	Reject Null Hypothesis
investor_name	24250.6042	0.0	Reject Null Hypothesis
investor_country_code	2089.3657	0.0	Reject Null Hypothesis
investor_region	4245.4457	0.0	Reject Null Hypothesis
investor_city	7175.2004	0.0	Reject Null Hypothesis
funding_round_type	16007.3625	0.0	Reject Null Hypothesis
funded_year	1154.331	0.0	Reject Null Hypothesis
raised_amount_usd	77929.0000	0.0	Reject Null Hypothesis

John W.Graham (2009) defined missing data as the data value that is not stored for a variable in the observation of the interest.. After feature selection process, all missing values from data were removed due to a significant effect on conclusions that can be drawn from the data. As a result, dataset has 77,929 rows and 13 attributes including target column Y.

As it has been seen from dataset, there is significant imbalance between majority of classes: variable “Yes” (89 %). and “No” (10%) in target column Y. Imbalanced data usually causes bias in classification and it leads to poor generalization in performance.

As statistical approach suggests, in order to address such class imbalance random resample the training dataset is needed. Liu, We and Zhou (2008) stated that the idea of under-sampling is to balance the data by filtering overrepresented data and have the same number of examples for each class. Random under- and over-sampling and combination of both methods had been applied to balance dataset. In addition, random under-sampling method removes rows of major classes while random over-sampling duplicates rows of minor classes. Table 8 below shows scatter plot of imbalanced classification dataset, scatter plot of imbalanced dataset with random under-, over-sampling and combination of both methods. These codes and results can be found at GitHub: [https://github.com/sharafjul/CIND-820-](https://github.com/sharafjul/CIND-820-Project/blob/main/Under_Sampling_with_RepeatedStratifiedKFold.ipynb)

[Project/blob/main/Under_Sampling_with_RepeatedStratifiedKFold.ipynb](https://github.com/sharafjul/CIND-820-Project/blob/main/Under_Sampling_with_RepeatedStratifiedKFold.ipynb)

Table 8. Scatter Plots with random under-sampling, over-sampling and combination of both methods

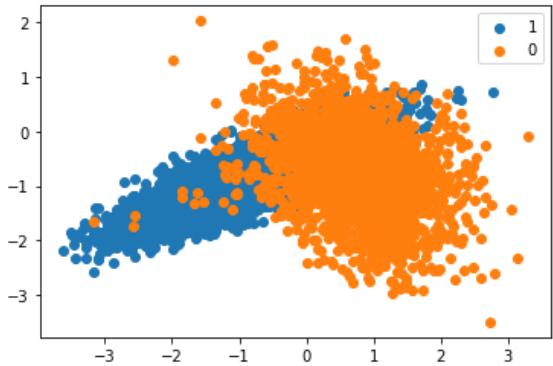
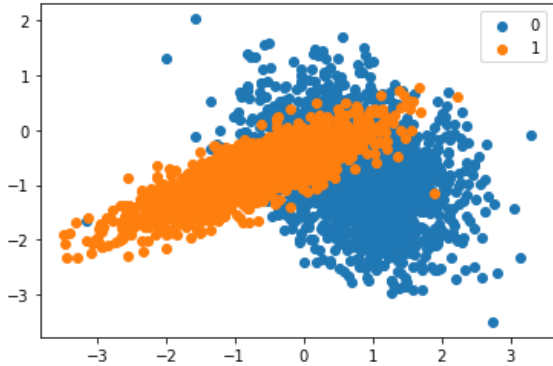
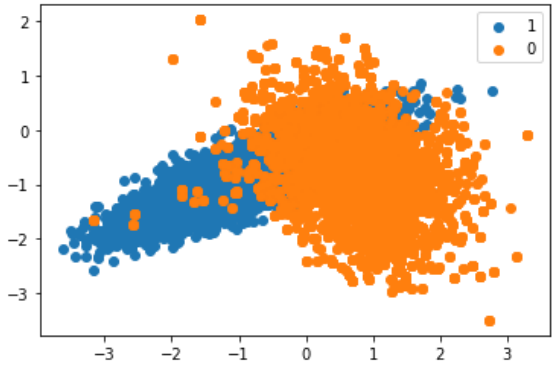
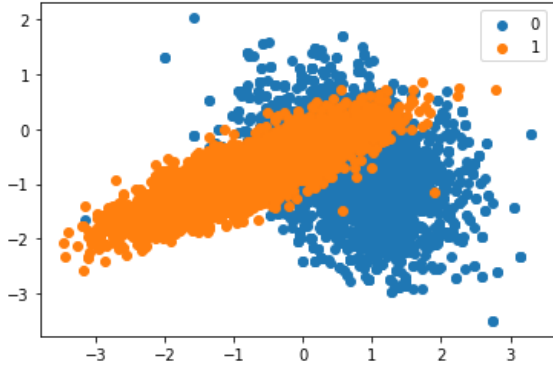
Name	Scatter Plot	Counter
Imbalanced classification dataset		1: 7977, 0: 2023
Resampling with RandomUnderSampler		0: 2023, 1: 2023
Resampling with RandomOverSampler		0: 7977, 1: 7977
Resampling with Combining Random Oversampling and Undersampling		0: 3988, 1: 4985

Table 9 summarized evaluation of each model results from the imbalanced dataset using under-, over-sampling methods and combination of both. Method Pipeline and Repeated Stratified K Fold cross validation were applied to evaluate the F1 score. The F1 score can be defined as a harmonic mean of the precision and recall, where the $F1 = 0$ is the worst possible score and $F1 = 1$ is a perfect score indicating that the model evaluates each observation correctly. According to Table 9, random over-sampling method has higher F1 in each model. This result sample will be applied on the next steps of data partitioning and in ML models. These codes and results can be found below at GitHub:

https://github.com/sharafjul/CIND-820-Project/blob/main/ML_DecisionTreeClassifier.ipynb

https://github.com/sharafjul/CIND-820-Project/blob/main/ML_RandomForestClassifier.ipynb

https://github.com/sharafjul/CIND-820-Project/blob/main/ML_Naive_Bayes_Classifier.ipynb

https://github.com/sharafjul/CIND-820-Project/blob/main/ML_KNeighborsClassifier.ipynb

Table 9. F1-score of data within Repeated Stratified K Fold cross validation

	Random UnderSampling	Random OverSampling	Combining Random Oversampling and Undersampling
Model	F1 Score		
Decision Tree Classifier	0.830	0.960	0.920
Random Forest Classifier	0.859	0.964	0.935
Naive Bayes Classifier	0.854	0.863	0.857
KNeighbours Classifier	0.853	0.905	0.877

Abadi (2009) defined data partitioning is the process of splitting data into different subsets for machine learning – training, validation, and testing. In this project, the data was split into training and testing subsets using Python Sklearn. Brownlee (2017) stressed that the purpose of training dataset is to train the model, while the test dataset is for the final evaluation of the model. In order to prevent sampling bias, Sklearn package randomly divided dataset into training and test sets, 70 % and 30% respectively.

Next, two classes were evaluated and converted to a binary classification. Classification is a supervised machine learning technique used to train models which predict classes based on the training from the training dataset. After that, the Decision Tree, Naïve Bayes, Random Forest and KNeighbours classification techniques were deployed to train, test and evaluate the performance of the dataset.

Table 10 shows results of each model: Decision Tree with 4 the maximum depth of the tree (max_depth= 7, 11, 15, none), RandomForestClassifier with 4 The maximum depth of the tree (max_depth= 7, 11, 15, none), Naive Bayes Classifier with Gaussian Naive Bayes algorithm and KNeighborsClassifier with 4 different number of neighbors (n_neighbors = 2, 5, 10, 15).

Gupta (2017) defined Decision Tree as a classification technique that uses attributes in a dataset to predict a class based on decisions. Furthermore, Decision trees are referred as classification for qualitative attributes or regression trees for continuous variables. Python Sklearn packages were applied to create the model. The Table 10 below summarizes the accuracy for decision trees with different values for max_depth, so decision tree with a max_depth = 7 has the lowest accuracy (88%) and max_depth = None (94%, highest accuracy). These codes and results can be found at GitHub:

https://github.com/sharafjul/CIND-820-Project/blob/main/ML_DecisionTreeClassifier.ipynb

Breiman (2001) defined the Random Forest classifier as a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Python Sklearn packages generated the RandomForestClassifier model. As input to this model 4 separate values of max_depth have been applied. Application of Random Forest model resulted in max_depth = None (the highest accuracy, 95%) and depth of the tree max_depth = 7 (the lowest accuracy, 88%). These codes and results can be found at GitHub:

https://github.com/sharafjul/CIND-820-Project/blob/main/ML_RandomForestClassifier.ipynb

Table 10. Evaluation Metrics for ML models

ML Model	Parameters	Confusion Matrix	Accuracy	ROC AU
Decision Tree Classifier	max_depth=7	[[2188 203] [367 2029]]	0.881	0.8810
	max_depth=11	[[2172 219] [245 2151]]	0.903	0.9031
	max_depth = 15	[[2281 110] [238 2158]]	0.927	0.9273
	max_depth = None	[[2346 45] [209 2187]]	0.947	0.9470
RandomForestClassifier	max_depth=7	[[2155 236] [310 2086]]	0.886	0.8860
	max_depth=11	[[2252 139] [241 2155]]	0.921	0.9206
	max_depth=15	[[2328 63] [200 2196]]	0.945	0.9451
	max_depth = None	[[2350 41] [182 2214]]	0.953	0.9534
Naive Bayes Classifier (GaussianNB ())	--	[[2100 291] [405 1991]]	0.855	0.8546
KNeighborsClassifier	n_neighbors=2	[[2367 24] [377 2019]]	0.916	0.9163
	n_neighbors=5	[[2235 156] [355 2041]]	0.893	0.8933
	n_neighbors=10	[[2162 229] [367 2029]]	0.875	0.8755
	n_neighbors=15	[[2128 263] [311 2085]]	0.880	0.8801

Harry Zhang (2004) defined Naive Bayes as one of the most efficient and effective inductive learning algorithms for machine learning and data mining. The Naive Bayes considers all attributes to be of equal importance as opposed to the decision tree in determining the more important attributes. The GaussianNB submodel type of the Naïve Bayes methods was used to build ML model. The accuracy of this model is 85.5%. These codes and results can be found at GitHub: https://github.com/sharafjul/CIND-820-Project/blob/main/ML_Naive_Bayes_Classifier_.ipynb

According to Goldberger, Roweis, Hinton and Salakhutdinov (2005) Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. Python Sklearn packages were used to create the KNeighborsClassifier model and 4 different number of N-neighbors was applied in this model. Model with `n_neighbors = 2` has higher accuracy (91%) than others. These codes and results can be found at GitHub: https://github.com/sharafjul/CIND-820-Project/blob/main/ML_KNeighborsClassifier.ipynb

References

- KPMG Global. (2022). Venture Pulse Report Q4'21. Global analysis of venture funding. <https://www.kpmg.us/insights/2022/venture-pulse-q4-2021.html>
- Lerner, J., & Tag, J. (2013). Institutions and Venture Capital. *Industrial and Corporate Change*, 22(1), 153-182. <https://doi.org/10.1093/icc/dts050>
- Allen, F. (2012). Trends in Financial Innovation and Their Welfare Impact: An Overview. *European Financial Management*, 18(4), 493-514. <https://doi.org/10.1111/j.1468-036x.2012.00658.x>
- Popov, A., & Roosenboom, P. (2013). Venture Capital and New Business Creation. *Journal of Banking & Finance*, 37(12), 4695-4710. <https://doi.org/10.1016/j.jbankfin.2013.08.010>
- Paul Gompers & Josh Lerner. (2001). The Venture Capital Revolution. *Journal of Economic Perspectives*, (15), 145–168
- William Janeway, Ramana Nanda and Matthew Rhodes-Kropf. (2021). Venture Capital Booms and Startup Financing. *Cambridge Working Papers in Economics* (14-16)
- A.V. Bruno, A.C. Cooper. (1982). Patterns of development and acquisitions for Silicon Valley startups *Technovation*, pp. 275-290. <https://www.sciencedirect.com/science/article/pii/0166497282900086>
- Juan B.Roure, Modesto A.Maidique. (1986). Linking prefunding factors and high-technology venture success: An exploratory study. *Journal of Business Venturing* Volume 1, Issue 3. 295-306. <https://www.sciencedirect.com/science/article/pii/0883902686900066>
- Oh, S., Jang, P., & Kwak, G. (2022). Enhancing the efficiency of governmental intervention in the venture capital market: The monitoring effect. *Economic Analysis and Policy*, 75, 450–463. <https://doi.org/10.1016/j.eap.2022.04.014>
- Barnes, R. (2021, August 9). What is an Outlier? Retrieved from The Data School: <https://dataschool.com/fundamentals-of-analysis/what-is-an-outlier/>
- Frost, J. (n.d.). 5 Ways to Find Outliers in Your Data. Retrieved from Statistics By Jim: <https://statisticsbyjim.com/basics/outliers/>
- Grace-Martin, K. (n.d.). Outliers: To Drop or Not to Drop. Retrieved from The Analysis Factor: <https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/>
- Alberto Quesada, A. (2017, January). 3 methods to deal with outliers. Retrieved from KD Nuggets: <https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.html>

Kaski, Samuel (1997) “Data exploration using self-organizing maps.” *Acta polytechnica scandinavica: Mathematics, computing and management in engineering series no. 82*.

Mitchell, Tom (1997). *Machine Learning*. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.

Kjell Johnson, Max Kuhn (May 2013) *Applied Predictive Modeling*, 488

John W. Graham (2009). Missing data analysis: making it work in the real world. *Annu Rev Psychol.*; 60:549–576.

Xu-Ying Liu; Jianxin Wu, Zhi-Hua Zhou (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2), 539-550.

Abadi, D. (2009). Data Partitioning. In: LIU, L., ÖZSU, M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_688

Brownlee, J. (2017, July 14). *Machine Learning Process: What is the Difference Between Test and Validation Datasets?* Retrieved from *Machine Learning Mastery*: <https://machinelearningmastery.com/difference-test-validation-datasets>

Gupta, P. (2017, May 17). *Decision Trees in Machine Learning*. Retrieved from *Towards Data Science*: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>

Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>

Zhang, Harry. (2004). *The Optimality of Naive Bayes*. Faculty of Computer Science. (PDF) <https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf>

J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov. (May 2005). Neighbourhood Components Analysis. *Advances in Neural Information Processing Systems*, Vol. 17, pp. 513-520. <https://cs.nyu.edu/~roweis/papers/ncanips.pdf>