Startups: A Success Prediction Model with Binary experiment model

Marwan Otrok[1] and Sugarbayar Enkhbayar[2]
[1] University of Warsaw, Faculty of Economics
m.otrok@student.uw.edu.pl
[2] University of Warsaw, Faculty of Economics
s.enkhbayar@student.uw.edu.pl

Abstract: This study analyzes the factors that influence the success of startups. It aims to develop a success versus failure prediction model regarding the global entrepreneurship ecosystem. Our study considers two categories that influence the success: characteristics of founders and characteristics of startups. In this study, data was obtained from a dataset that was offered in a previous competition by CrowdAnaltytix, with the objective to identify factors that determine the success of startups in the Big Data space. The features that are included in this dataset are aligned with the two categories that were pre-selected from existing literature review with the aim to further verify its impact on startup success. The empirical results show that previous work experience of founders obtained from medium-sized and large companies have a significant influence in startups success, unlike the education level of founders that interestingly is found to have significant impact on startup success. Startups with … are more likely to be unsuccessful cases. Meanwhile, the type of business venture in the Big Data space, as a startup characteristic, is significant and positively related to the success of startups. Overall, the success and failure prediction model presents an ability to accurately predict a specific startup as success of 69%.

Keywords: Startup, Logit model, Probit model, Marginal effect, KNN model

1. Introduction

Startups are a driving force for economic development and can bring substantial profits to venture capital firms if they succeed. Being able to predict the success of startups gives investors a significant advantage over their competitors.

Venture capital (VC) funds represent a type of private equity investment specifically targeted at early-stage startups displaying high growth potential. These funds are provided by venture capitalists who inject capital into these startups in exchange for an ownership stake, with the ultimate goal of realizing significant returns on their investments over a prolonged period. Venture capital firms possess specialized expertise and experience in assessing startup prospects and offering strategic guidance to their portfolio companies. As elucidated by Ries (2011), startups emerge as entities established under conditions of great uncertainty, endeavoring to develop innovative products or services. It is widely acknowledged that a substantial proportion of startups fail, with an estimated failure rate of approximately 90%. Even among startups that secure venture capital backing, over 75% either fail outright or grapple with survival challenges (Picken, 2017). Consequently, the ability to forecast startup success assumes paramount importance for venture capitalists as it enables them to optimize investment decisions, allocate resources judiciously, provide valuable strategic counsel, and preserve their reputation as accomplished investors. By effectively evaluating the potential of startups, venture capitalists can propel innovation, stimulate economic growth, and yield substantial returns for their investors. Although VC funds typically offer favorable returns on investment, studies indicate that during the 2000s, they exhibited lower performance relative to the S&P 500 index (Harris et al., 2014). Consequently, venture capital funds face the formidable task of identifying startups with a greater likelihood of achieving success in order to overcome this challenge.

Over the past few decades, researchers have extensively researched and investigated factors that contribute to venture success or failure. They have explored various models, which analyzed a wide range of variables (financial and nonfinancial predictors) to explain the outcomes of startups worldwide.

Our paper aims to examine and identify the drivers for startup success, by leveraging building upon the existing literature for categorizing the factors contributing to business success. Lussier (1995) employed logistic regression analysis to anticipate the performance of young companies within the United States. The research methodology involved the collection of survey data and the inclusion of fifteen explanatory variables, which encompassed diverse aspects such as capital accessibility, record keeping and financial controls, industry and management experience, planning, professional advisory support, educational background, staffing, product or service timing, economic conditions, age of the owner, presence of partners, prior business ownership by parents, minority status, and marketing skills.

This widely recognized model has garnered significant popularity and adoption as a forecasting tool for both success and failure rates across various industries, companies of different sizes, and within six different countries over the span of the past two decades. Notably, the model exhibited a predictive capacity ranging from 63% to 85%, indicating its effectiveness in anticipating business outcomes.

By leveraging Lussier's comprehensive framework and utilizing logistic regression analysis, our paper intends to extend the understanding of business success factors in a specific context, allowing for the exploration of additional variables or modifications to the existing model. Through this research endeavor, we seek to contribute to the existing body of knowledge in the field and provide insights that

can inform decision-making processes and enhance the prospects of venture capital and entrepreneurial ventures.

2. Literature Review

2.1. Startups

Defining startups is a complex task due to the evolving nature of entrepreneurial ventures. Various scholars and organizations have proposed different definitions to capture the essence of startups. According to Blank and Dorf (2012), a startup is a temporary organization designed to search for a repeatable and scalable business model in an environment of extreme uncertainty. Ries (2011) defines a startup as a human institution designed to deliver a new product or service under conditions of extreme uncertainty.

Startups are typically focused on discovering innovative solutions, refining their products or services, and exploring market opportunities. They are driven by a desire to address unmet needs in the market, challenge existing industry norms, and create value through unique value propositions. Startups often strive to disrupt established markets by leveraging technological advancements or offering novel approaches to problem-solving. Startups often seek external funding, such as venture capital, to fuel their growth and expand their market presence. Startups aim to develop business models that can be replicated and scaled up quickly, allowing them to capture larger market shares and generate substantial returns on investment. Startups are typically founded by visionary entrepreneurs who are driven by a passion for their ideas, a desire for autonomy, and a willingness to take calculated risks. Entrepreneurial intent is a key distinguishing factor between startups and other types of organizations, highlighting the proactive and innovative mindset that underpins the startup culture defines startups as newly formed organizations that introduce innovative products or services with the intention of achieving rapid growth.

It is also important to stress that startups are not smaller versions of large enterprises. Steve Blank (2010) explains the main organizational difference between startups and large enterprises and its implications for innovation activities. Because corporations develop strategies and design structures in the sense that their proven business models can efficiently function, innovation within such organizations encounter more obstacles when compared to a startup organizational structure, as startups tend to be more flexible due to having less structural and strategic constraints (Blank, 2014).

2.2. Startup Success

Identifying and measuring startup success can be difficult because it is a relative measure. Startup success is a multifaceted concept that goes beyond mere financial performance. It encompasses various dimensions, including financial outcomes, strategic achievements, and socio-economic impact. Success can be measured in different ways and it will depend on the enterprise goals which can be financial or non-financial, simple pre-defined expectations or founders' behavior. In 1986, Barney (1986) defined success as a measure of performance that happens when the business creates value for its customers in a sustainable and economically efficient manner. Other measures of performance can be employment growth, revenues, profit and other financial performance measures (MayerHaung et al., 2013).

Zbikowski and Antosiuk (2021) conducted a study in which they sourced their data from Crunchbase, a platform that provides comprehensive information on startups. The dataset included

various features, such as initial public offerings (IPOs), acquisitions, and investment rounds, which were utilized to assess the success of startups. Specifically, the researchers defined success based on whether a startup had undergone one of the following outcomes: an IPO, acquisition by another company, or securing a second investment round. To identify startups that experienced failure, they employed an indicator based on the response from the companies' homepage URLs, considering an inactive homepage as a reliable marker of a startup's failure in the market. The incorporation of such information proved valuable in establishing a target variable for assessing a company's success.

2.3. Deterministics of Startup Success

Over the past few decades, various studies have been conducted in order to further explore and comprehend the prediction of startup success and failure along with their performance evaluation, but there is still no widely acknowledged list of predictors that would impact startup success. Multiple explanatory variables for startup success/failure were taken into consideration in various studies, which were later grouped into different categories by different researchers. Lussier (1995) and Carter and Auken (2006) categorized the success factors in four categories: characteristics of the founders and characteristics of the startup. In our paper, two categories will be selected to further investigate their impact on startup success, and accordingly our primary and secondary hypotheses shall be formulated - one for each category. The three categories are further discussed as followed:

2.3.1. Characteristics of the founders

Founders play a crucial role in startups, and their characteristics shape the startup culture and its interaction with the business environment. Human capital, encompassing experience, knowledge, age, and education, is recognized as a critical factor for organizational performance. It is positively correlated with founders' abilities to identify and capitalize on business opportunities, develop effective plans and strategies, and acquire resources such as financial and physical capital.

2.3.1.1 Education / Academic Experience

The influence of formal education on entrepreneurial success deserves attention. Colombo and Grilli (2005) conducted a study revealing a positive association between the total years of education and the likelihood of securing capital. However, they found no significant impact of education on venture growth. Notably, they observed a positive correlation between education in economics, management, and technical/scientific fields and venture growth. Building upon this, Hsu (2007) delved deeper into the relationship and identified a negative correlation between founders holding a PhD and receiving venture capital in established industries, but a positive correlation in emerging industries.

The significance of education extends beyond fundraising. Kalyanasundaram et al. (2021) conducted research that corroborated the positive effect of higher education on the survival rates of startups. The study emphasized that higher levels of education within the entrepreneurial team contribute to their proficiency in technology creation and diffusion. Furthermore, elevated education levels indicate a higher degree of maturity in dealing with adversities and contingencies. The researchers argued that the knowledge and skills acquired through higher education programs provide individuals with the necessary tools to navigate the complexities and challenges associated with establishing and

sustaining a startup venture. Consequently, teams with a strong educational foundation are better equipped to overcome hurdles and demonstrate a higher likelihood of long-term survival and success.

2.3.1.1 Professional Experience

Founders' experience, skills, and industry-specific knowledge contribute significantly to entrepreneurial talent and performance. Management experience, marketing skills, and the understanding of products, processes, and technology also have positive influences on business success (Santos da Silva et al., 2016).

Spanjer et al. (2017) finds that experience deteriorates over time resulting in older experience relating to decreased entrepreneurial performance while recent experience increases performance. This is explained by entrepreneurs with a less recent variety of experience potentially having outdated information or an inaccurate memory from the experience leading to incorrect conclusions.

Industry-specific knowledge has been shown to increase the likelihood of success, as know-how reduces the liability of newness, decreasing the risk of failure for startups. Cassar (2014) explains this to be due to superior forecasting abilities that come with industry experience. On the other hand, there are other studies that argue there is no empirical evidence that industry-specific knowledge has a positive impact on venture growth, which in turn can be contradicting the current research consensus (Marino and Noble, 1997).

Prior startup experience has also been shown by various studies to have a positive impact on startups increasing their likelihood of success, especially, if the previously established startups were successful (Colombo and Grilli, 2005). Whereas, Marino and Noble (1997) did not find empirical evidence that previous start-up experience significantly influenced the survival and growth of a startup.

2.3.2. Startup Characteristics

The characteristics of startups represent a distinct category that exerts influence on the success of a business venture. Firstly, the skills and knowledge possessed by the startup team play a crucial role in determining its success. The size of the founder team, in particular, emerges as a factor capable of impacting business outcomes by facilitating the accumulation of entrepreneurial talent. When founders with complementary competencies and skills join forces, the cognitive and managerial capacity of individual founders expands. While the positive influence of team size on performance has been acknowledged, it is important to note that larger team sizes do not automatically guarantee better performance. The coordination and communication challenges inherent in managing a larger team can impede effectiveness (Brinckmann & Högl, 2011; Mayer-Haung et al., 2013).

The presence of professional advisors has also been identified as a contributing factor to startup success. Lussier and Marom (2014) demonstrate that professional advisors facilitate access to information networks, which provide specific and valuable data resources. The act of seeking information itself reflects comprehensive planning and a higher level of managerial sophistication. Thus, the presence of professional advisors increases the likelihood of startup success by enhancing access to critical resources and expertise.

Furthermore, the relationship between product or service timing and business success has received attention in research. Startups that introduce extremely new and innovative products or services, as well as those that offer products or services that have become outdated, face a higher

likelihood of survival compared to ventures operating in the growth stage (Lussier et al., 2016). The timing of product or service delivery can significantly impact the competitive landscape and consumer demand, ultimately influencing the success and longevity of a startup.

In summary, the characteristics of startups, including the skills and knowledge of the team, the presence of professional advisors, and the timing of product or service offerings, collectively contribute to the success and sustainability of a business venture. These factors must be carefully considered and managed to optimize the prospects of startup success.

3. Hypotheses development

3.1. Characteristics of the founders

The characteristics of founders significantly impact the success of startups. Human capital, including experience, education, and industry-specific knowledge, plays a crucial role in discovering opportunities, acquiring resources, and formulating effective strategies. Additionally, management experience, marketing skills, and industry know-how contribute positively to business success. Taking into consideration the above factors, the primary hypothesis proposed on this study is as followed:

H1: The characteristics of a founder have a significant impact on startup success.

3.2. Characteristics of Startups:

The characteristics of the startup play an essential role in determining its success. The size of the founder team is an important factor, as it allows for the accumulation of entrepreneurial talent and complementary competencies. Additionally, the presence of professional advisors and the ability to attract and retain quality employees contribute to release also have positive influences on startup success. Therefore, it is hypothesized that the business succeeds by providing access to information networks and increasing managerial sophistication. The existence of a specific business plan and the timing of product or service characteristics of the startup significantly influence its success.

H2: The startup characteristics have a significant influence in startup success

4. Methodology

4.1 Linear probability model, logit model and probit model

When dealing with a dichotomous dependent variable, such as in our case "success" or "failure," of startups in the Big Data space, logistic regression (logit) is generally preferred over linear probability in most cases for several reasons. Firstly, logistic regression models the relationship in terms of odds and allows for non-linear relationships, ensuring predicted probabilities are always within the valid range. Secondly, logistic regression is considered to be more robust to heteroscedasticity, and lastly, the coefficients (log odds ratios) can be directly interpreted in terms of the odds of the event occurring. This makes it easier to understand and communicate the impact of the independent variables on the probability of the outcome, whereas the coefficients in the linear probability model represent the change in probability, which may not be as intuitive.

In binary choice models we try to explain probability of success with explanatory variables. If the dependent variable is success, it takes 1. If the dependent variable is a failure, it takes 0.

The Linear Probability Model

This is a general case of linear regression, in which we model the expected value of Y as a linear function of independent variables.

$$E(Y) = XB$$

$$E(Y) = 1[Pr(Y = 1)] + 0[Pr(Y = 0)] = Pr(Y = 1)$$

So we model:

$$Pr(Y_i = 1) = X_i B + e$$

That is, we just do a linear OLS regression of Y on X. This is the linear probability model (LPM). It amounts to fitting regular OLS regression to a binary response variable. But there are many problems with LPM.

- e is heteroscedastic
- We cannot constraint XB to the 0-1 interval
- The LPM's estimators are biased and inconsistent

So instead of LPM, we will derive the models for logit and probit models.

Logit and Probit model

The latent variable approach treats dichotomous dependent variables as essentially a problem of measurement.

- That is, there exists a continuous underlying variable; we just haven't measured it.
- Instead, we have a dichotomous indicator of that underlying latent variable.

Call this latent variable Y^*.

The underlying model is then:

$$Y_i^* = X_i B + u_i$$

This model has the usual OLS-type assumptions; in particular, that u_i is distributed according to some symmetrical distribution.

However, we observe only the following realizations of Y^*:

$$Y = 0 \; if \; Y_i^* < 0$$

$$Y = 1 \; if \; Y_i^* >= 0$$

So, we can write:

$$Pr(Y_i = 1) = Pr(Y_i^* >= 0) = Pr(u_i <= X_i B)$$

Logit

If we assume that the u_i follow a standard logistic distribution, we get a logit model.

- The standard logistic probability density is:

$$Pr(u) = \lambda(u) = exp(u)/(1 + exp(u))^2$$

We can further write this now, as:

$$\Pr(Y_i = 1) \equiv \Lambda(\mathbf{X}_i\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i\boldsymbol{\beta})}$$

This is the basic form of the probability for the logit model. To get a probability statement for every observation in our data, we want to think of the probability of getting a zero (one) given the values of the covariates and the parameters. That is, the likelihood for a given observation i is:

$$L_i = \left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{Y_i}\left[1-\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]^{1-Y_i}$$

That is, observations with Y = 1 contribute $Pr(Y_i = 1|X_i$to the likelihood, while those for which Y = 0 contribute $Pr(Y_i = 0|X_i$. Assuming that the observations are conditionally independent, we can take the product over the N observations in our data to get the overall likelihood:

$$L = \prod_{i=1}^{N}\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)^{Y_i}\left[1-\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]^{1-Y_i}$$

Taking the natural logarithm of this yields:

$$\ln L = \sum_{i=1}^{N} Y_i \ln\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right) + (1-Y_i)\ln\left[1-\left(\frac{\exp(\mathbf{X}_i\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i\boldsymbol{\beta})}\right)\right]$$

We can then maximize the log-likelihood with respect to the $B\hat{}s$ to obtain our MLEs, in the manner in which we discussed last week.

The probit

In the probit model we use normal distribution. The probability setup of the probit thus looks like:

$$\begin{aligned}\Pr(Y_i = 1) &= \Phi(\mathbf{X}_i\boldsymbol{\beta}) \\ &= \int_{-\infty}^{\mathbf{X}_i\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\mathbf{X}_i\boldsymbol{\beta})^2}{2}\right) d\mathbf{X}_i\boldsymbol{\beta}\end{aligned}$$

And the corresponding log-likelihood is:

$$\ln L = \sum_{i=1}^{N} Y_i \ln\Phi(\mathbf{X}_i\boldsymbol{\beta}) + (1-Y_i)\ln\Phi(\mathbf{X}_i\boldsymbol{\beta})$$

Odds ratio

Also, we can calculate the odds ratio using estimates of variables.

$$\Omega \equiv \text{Odds(Event)} = \frac{\Pr(\text{Event})}{1-\Pr(\text{Event})}.$$

Interpretation: We can interpret only qualitatively. It means that we can interpret only their signs.

Marginal effect: If we are interested in quantitative effects on variables. We should use marginal effects. We obtain marginal effects for average characteristics.

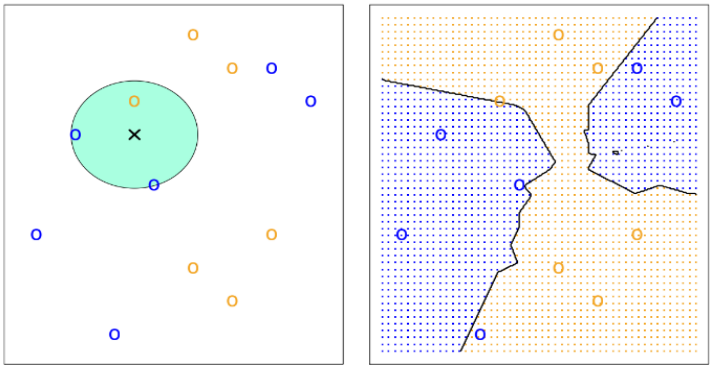
R squares: We have four R squared statistics in the logit model and explanations are different. For example: McFadded R2, McKelvey-Zavoina R2, Count R2, and Adjusted Count R2.

4.2 K Nearest Neighbor model

KNN model is classification model. for the natural number K and the observation from the test sample x_0, the KNN classifier identifies K points from the training sample, which are located closest to x_0–

its nearest neighbors. The new observation is classified into the group most represented among K neighbors.

Figure: K-nearest neighbor example



The KNN algorithm treats each variable as a separate dimension of space - taking into account p variables, we operate in the p-dimensional space. There are many ways to measure the distance of objects. The most common method in the KNN method is the Euclidean distance - the length of the shortest segment connecting two points. It is calculated as the square root of the sum of squares of differences corresponding to the coordinates of individual points. For points i and j and p variables the Euclidean distance can be calculated as. Also, we can use city distance and minkowski distance.

$$d_e(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \ldots + (x_{pi} - x_{pj})^2}$$

The choice of the value of K parameter has a key impact on the classification results. For k=1 the method is most flexible, it fits very closely to the data - in such cases it has large variance and low bias. With the increase of K the flexibility of the method decreases, and the boundaries between groups become more and more linear.

5. Dataset

5.1 Interpretation of the data

Initially, data was collected from Crunchbase, one of the largest platforms providing business information on startups worldwide. This data will primarily be utilized for a future thesis project. The intention was to begin by selecting the necessary variables from the various tables in Crunchbase and potentially augment the dataset with founder characteristic information from LinkedIn. However, due to time constraints, this process proved to be lengthy. Consequently,alternative datasets were sought that contained adequate data and features to validate the formulated hypothesis. Fortunately, a dataset dedicated to a previous competition organized by CrowdAnalytix was obtained, which aimed to identify factors determining startup success in the Big Data and Analytics space. The features included in this dataset align with the four pre-selected categories derived from existing literature reviews, allowing for further examination of their impact on startup success. These categories encompass founder characteristics, startup characteristics, capital, and external factors.

Our data contains information of 472 startups in 2014. Initial data has 116 variables, so we decided to use the following variables in our analysis. Then we categorized these variables into 4 categories such as founder characteristics, startup characteristics, capital, and external factors.

Table: Initial selected variables description

Groups	Variable	Description
Dependent variable	Company status {1, 0}	1 - successful startup, 0 - failed startup
Founder characteristics	Worked.in.top.companies {Yes, No}	Yes - founder worked in top companies No - founder didn't work in top companies
	Average.size.of.companies.worked.for.in.the.past {large, medium, small}	Size of company founder worked in the past
	Have.been.part.of.startups.in.the.past {Yes, No}	Was the startup part of another startup in the past?
	Have.been.part.of.successful.startups.in.the.past {Yes, No}	Was the startup part of another successful startup in the past?
	Consulting.experience {Yes, No}	Yes - Founder has consulting experience No - Founder hasn't consulting experience
	Was.he.or.she.partner.in.Big.5.consulting.{Yes, No}	Is she/he a partner in big 5 consulting companies?

Table: Initial selected variables description - continue

Groups	Variable	Description
Startup characteristics	Age.of.company.in.years {numeric}	the period since the establishment of the company
	Focus.on.private.or.public.data {Yes, No)	Does the company focus on private or public data?
	Focus.on.consumer.data {Yes, No}	Does the company focus on consumer data?
	Focus.on.structured.or.unstructured.data {structured, unstructured}	Companies focus on structured or unstructured data.
	Machine.Learning.based.business {Yes, No}	Is the company a business based on machine learning? Or not?
	Predictive.Analytics.business {Yes, No}	Is the company a business based on predictive analytics? Or not?
	Speech.analytics.business {Yes, No}	Is the company a business based on speech analytics? Or not?
	Prescriptive.analytics.business {Yes, No}	Is the company a business based on prescriptive analytics? Or not?
	Big.Data.Business {Yes, No}	Is the company a business based on big data? Or not?
	Product.or.service.company. {Product, Service, Both}	Does the company focus on product, service, or both?
	Local.or.global.player {local, global}	Is the company local or global?
	Number.of.Investors.in.Seed {numeric}	Number of investors
	Internet.Activity.Score {numeric}	Higher score is more activity on internet
	Number.of.Co.founders {numeric}	Number of co-founder
	Number.of.of.advisors {numeric}	Number of advisors
	Team.size.all.employees {numeric}	Number of employees
	Number.of..Sales.Support.material {high, medium, low}	Size of sales support material
	Number.of..of.Partners.of.company {None, few, many}	Number of partners

5.2 Data cleaning and Initial variable selection

Data recoding: Data is survey data, so it requires a lot of data cleaning. We have done the following for it.

- Converted all character variable into lower letter
- Remove white spaces from all character variables
- Replaced non allowed (data recode) values to most frequently value

Data type: After that, we changed the character variable into factor variable. Also, we applied ordered factor type for following variables: number of sales support material, average size of companies worked in the past, and number of partners. Because these variables can be ordered from lowest to highest.

Missing value: Mainly, numeric variables have missing value. So we filled in missing values of numeric variables with median. Also, we filled in missing values of the character variable with the most frequent value.

Table: Number of missing values and filling method

Variables	Number of missing values	Fill in
Last funding amount	160	Median value
Team size all employees	68	Median value
Internet Activity score	65	Median value
Age of company in years	59	Median value
Number of investors in seed	49	Median value
Number of sales support material	48	Most frequently value

Outliers: We used the Interquartile method to find outliers of all numeric variables. We have a small dataset, so we replaced all outliers with median value.

Table: Number of outliers and filling method

Variables	Number of outlier	Fill in
Age of company	11	Median
Last funding amount	69	Median
Number of investors	79	Median
Internet activity score	55	Median
Number of co.founder	8	Median
Number of advisor	64	Median
Team size all employees	32	Median

6. Result

In order to test for non-linear relationships, we created the following two new variables using powers.

$$team.size.all.employees2 = team.size.all.employee * team.size.all.employe \quad (1)$$

$$age.of.company.in.years2 = age.of.company.in.years * age.of.company.in.years \quad (2)$$

6.1 Variable selection based on distribution

Before training the model, we checked the distribution of numeric explanatory variables. It gives us a lot of information about variable selection. We discovered the following variables are too right skewed. So we removed these variables from our model. Because these variables contain mostly one value, it is not necessary to use them in the analysis.

We also checked the frequency of categorical variables using a bar chart. 80% of the following 4 variables consist of only one value such as worked in top companies, was he or she a partner in big 5 consulting, speech analytics business, and number of partners company. So we decided to remove these

4 variables from our models. Generally we removed 6 variables from our initial variables based on their distribution.

Figure: Distribution of numerical variables

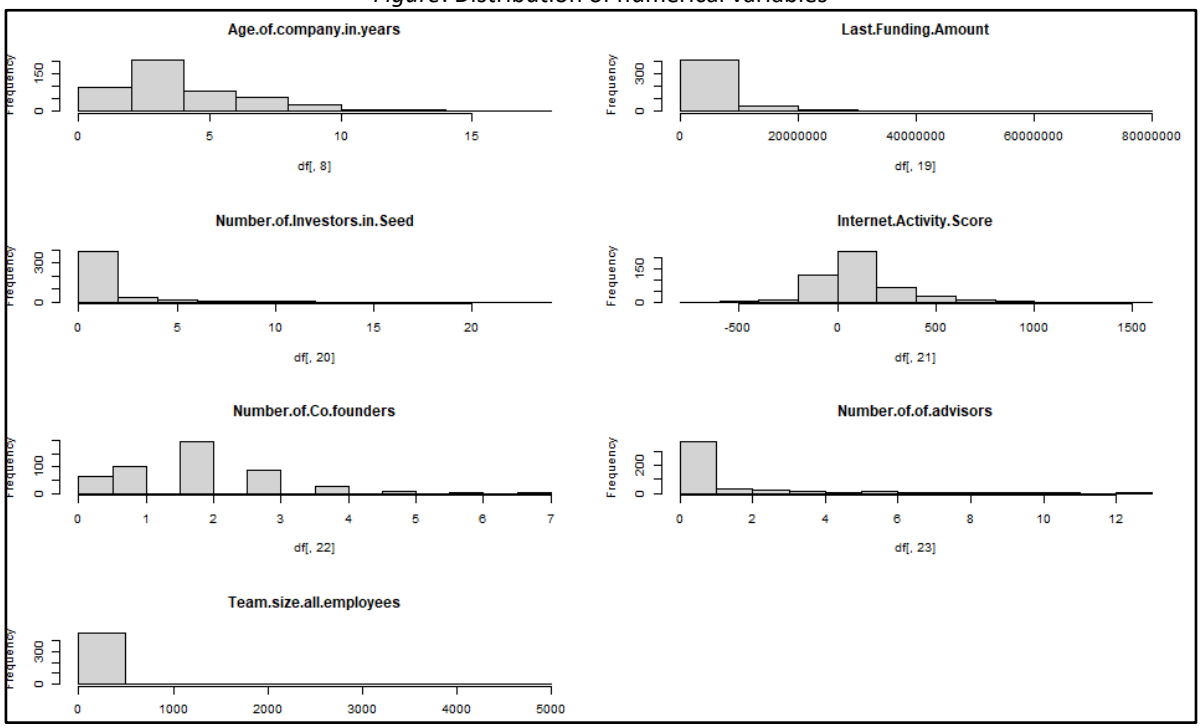
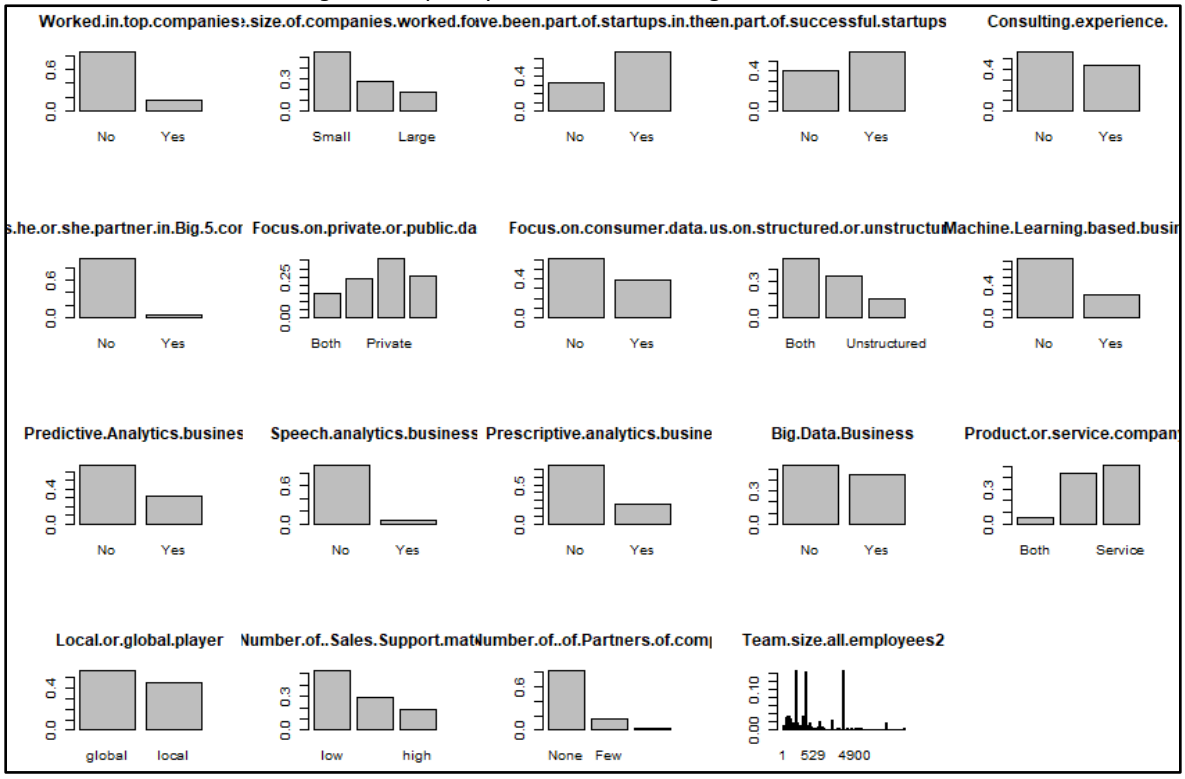


Figure: Frequency distribution of categorical variables



6.2 Descriptive analysis

We checked correlation between numerical variables. There is no strong correlation.

Table: Correlation between numerical variables

	Age.of.company.in.years	last.funding.amount	Internet.Activity.Score	Number.of.co.founders	Team.size.all.employees
Age.of.company.in.years	1	0.23	-0.25	-0.11	0.19
last.funding.amount	-	1	0.05	0.04	0.2
Internet.Activity.Score	-	-	1	0.12	0.1
Number.of.co.founders	-	-	-	1	0.05
Team.size.all.employees	-	-	-	-	1

Data split: We divided our data into train (70%) and test sets (30%). We train models on only train set. And we find the best model based on the performance of the model on the test set.

Table: Description of train and test sets

Type	0 (dependent variable - fail)	1 (dependent variable - success)
Train	117	214
Test	50	91

Multiclass explanatory variable: We can use variables with 2 options as explanatory variables in binary experiment model. But we can't use variables with more than 2 options. So we need to create a new dummy variable for each option.

Table: Recoding multiclass explanatory variables

Old variable	New dummy variables	Description
Average.size.of.companies.worked.for.in.the.past	average_L {0, 1}	1 - founder worked in large company
	average_M {0, 1}	1 - founder worked in medium company
	average_S {0, 1}	1 - founder worked in small company
Focus.on.structured.or.unstructured.data	struct_S {0, 1}	1 - focus on structured data
	struct_U {0, 1}	1 - focus on unstructured data
	struct_B {0, 1}	1 - focus on both

6.3 Logit, probit and linear probability model

Firstly, We used a backward variable selection method with logit and probit models to find the most important variables. Backward - begins with a model that contains all variables (full model). Then it starts removing the least significant variables one after another. We had 26 explanatory variables. The following 14 variables were identified as highly significant variables using the backward method. We used only these 14 variables in our further analysis.

We trained logit, probit, and LPM models using the most important 14 variables.

- We can't use LPM models with binary target variables. Because it applies a simple regression model for dummy dependent variables. When we use LPM, we will obtain heteroscedastic, biased and inconsistent. So we immediately removed this model
- Logit and probit models are very similar. They use different cumulative distribution functions. From the table below, we can see that the logit model is better than probit. Because the logit model has lower AIC than the probit model
- We checked homoscedastic of each model with breusch-pagan-test. P values of all models are lower than 0.05. So it means that these models are heteroscedastic. Because we can't reject the null hypothesis.

Table: Result of three models. Numbers shows estimates

Variable	Logit	Probit	LPM
(Intercept)	2.2.	1.13.	1.67***
Have.been.part.of.successful.startups.in.the.past.Yes	-0.76*	-0.52*	-0.06
Consulting.experience.Yes	0.61	0.4.	0.03
Age.of.company.in.years	-1.11*	-0.5*	-0.09*
Machine.Learning.based.businessYes	1.23*	0.4.	0.07
Predictive.Analytics.businessYes	0.99.	0.81**	0.11*
Prescriptive.analytics.businessYes	-1.09*	0.44	-0.06
Big.Data.BusinessYes	2.22***	-0.69*	0.23***
Local.or.global.playerlocal	-1.81***	1.09***	-0.22***
Internet.Activity.Score	0.005**	-1.02***	0.0007***
Age.of.company.in.year 2	0.07*	0.003**	0.006
average_L	1.21*	0.03.	0.13*
average_M	1.78***	0.62*	0.18***
struct_S	2.18***	1.05***	0.19***
struct_U	0.98.	1.06***	0.14*
AIC	**238.2**	241.01	~

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table: Breusch-Pagan of three models

	Logit	Probit	LPM
Breusch-Pagan test (p value)	0.01243	0.01243	0.01243

From the previous table, we determined the Logit model is the best model. Therefore, the Logit model will be described in more detail and additional tests will be tested.

Interpretation of coefficients: We can't explain the coefficient of the logit model quantitatively . But we can explain it qualitatively.

Positive relationship:

- If a startup is a **machine learning business**, the probability of success will be **increased**
- If a startup is a **predictive analytics business**, the probability of success will be **increased**

- If a startup is a **big data business**, the probability of success will be **increased**
- If a startup's **internet activity score** increase, probability of success will be **increased**
- If founder worked in **large company** in the past, probability of success will be **increased**
- If founder worked in **medium company** in the past, probability of success will be **increased**
- If startup focus on **structured** data, the probability of success will be **increased**
- If startup focus in **unstructured** data, the probability of success will **increased**

Negative relationship:

- If a startup has been **part of a successful startup** in the past, the probability of success will be **decreased**
- If a startup is a **prescriptive analytics business**, the probability of success will be **decreased**
- If a startup is a **local** player, the probability of success will be **decreased**

Non-linear relationship:

- Additional age of year will decrease the probability of success in the short term. But an additional age of year will increase the probability of success in the long term. It is a non-linear relationship.

Marginal effects: We can explain the marginal effect of logit model quantitatively.

Positive relationship:

- If a startup with **average characteristics** is a **machine learning** business, the probability of success will **increase by 13 percentage points**
- If a startup with **average characteristics** is **predictive analytics**, the probability of success will **increase by 11 percentage points**
- If a startup with **average characteristics** is **predictive analytics**, the probability of success will i**ncrease by 11 percentage points**
- If a startup with **average characteristics** is a **big data business**, the probability of success will **increase by 28 percentage points**
- If a startup's **internet activity score** with **average characteristics** increases by 1 unit, the probability of success will **increase by 0.07 percentage points**
- If a founder of startup with **average characteristics** worked in large company in the past, the probability of success will **increase by 0.12 percentage points**
- If a founder of startup with **average characteristics** worked in larmedium company in the past, the probability of success will **increase by 0.18 percentage points**
- If a startup with **average characteristics** focus on structured data, the probability of success will **increase by 24 percentage points**
- If a startup with **average characteristics** focus on unstructured data, the probability of success will **increase by 10 percentage points**

Negative relationship:

- If a founder with **average characteristics** has been **part of a successful startup** in the past, the probability of success will **decrease by 9 percentage points**
- If a startup with **average characteristics** is **prescriptive analytics**, the probability of success will **decrease by 17 percentage points**
- If a startup with **average characteristics** is a **local player**, the probability of success will **decrease by 26 percentage points**

Non-linear relationship:

- If a startup's age increases by 1 unit with average characteristics, it will decrease the probability of success by 14 percentage points in the short term. But If a startup's age increases by 1 unit with average characteristics, it will increase the probability of success by 1 percentage point.

Table: Marginal effect of logit model. Numbers shows estimates

Variable	Logit
Have.been.part.of.successful.startups.in.the.past.Yes	-0.09*
Consulting.experience.Yes	0.08
Age.of.company.in.years	-0.14**
Machine.Learning.based.businessYes	0.13*
Predictive.Analytics.businessYes	0.11*
Prescriptive.analytics.businessYes	-0.17.
Big.Data.BusinessYes	0.28***
Local.or.global.playerlocal	-0.26***
Internet.Activity.Score	0.0007**
Age.of.company.in.year 2	0.01*
average_L	0.12**
average_M	0.18***
struct_S	0.24***
struct_U	0.10*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Odds Ratio: We can compare the effect of 2 variables using odds ratio. We compared the following variables.

- 2.18/0.98=2.22 -> It means startup focus on structured data increases probability of success is the same way as 2.22 startup focus on unstructured data
- 1.78/1.21=1.47 -> It means founder of startup worked in medium company increases probability of success is the same way as 1.47 founder of startup worked in a large company
- 2.22/1.23=1.8 -> It means startup big data business increases probability of success is the same way as 1.8 startup machine learning business

R squares: We can measure different R squares. They are interpreted differently.

- McKelveyZavoina R squared is 0.75. It means that if a latent variable was observed then our model would explain 75 percent of its variation. It is quite similar to R2 in simple regression
- Count R squared is 0.88. It means that our model correctly predicts 88 percent of all observations
- Adjusted Count R squared is 0.67. It means that only 67 percent of all predictions were correct because of the variation of dependent variable

Table: Description of R squares

McKelveyZavoina	Count R2	Adjusted Count R2
0.75	0.88	0.67

Linktest: We used a simple linear model. And we wanted to check if our model is good enough. yhat variable is statistically significant. And yhat squared is not significant. We can state that our model is correct. It means that our model is enough to explain dependent variables.

Table: Result of linktest

```
glm(formula = y ~ yhat + yhat2, family = binomial(link = model$family$link))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0890  -0.4203   0.1655   0.3947   2.7956

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.11418    0.20105   0.568    0.570
yhat         1.06831    0.13135   8.133 4.19e-16 ***
yhat2       -0.04749    0.03897  -1.219    0.223
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 430.02  on 330  degrees of freedom
Residual deviance: 206.92  on 328  degrees of freedom
AIC: 212.92

Number of Fisher Scoring iterations: 7
```

Likelihood Ratio test with One explanatory variable(Age.of.company.in.year 2): We used likelihood ratio test to verify hypothesis testing. Because the logit model is estimated using the maximum likelihood method. We verify the following hypothesis.

H0: Age.of.company.in.year 2 = 0

The P value of the LR test is lower than 0.05. So we can reject the null hypothesis and the coefficient of Age.of.company.in.year.2 variables is not equal to zero.

Table: Result of likelihood Ratio test of logit model with one coefficient

```
Likelihood ratio test

Model 1: Dependent.Company.Status ~ Have.been.part.of.successful.startups.in.the.past. +
    Consulting.experience. + Age.of.company.in.years + Machine.Learning.based.business +
    Predictive.Analytics.business + Prescriptive.analytics.business +
    Big.Data.Business + Local.or.global.player + Internet.Activity.Score +
    Age.of.company.in.years2 + average_L + average_M + struct_S +
    struct_U
Model 2: Dependent.Company.Status ~ Have.been.part.of.successful.startups.in.the.past. +
    Consulting.experience. + Age.of.company.in.years + Machine.Learning.based.business +
    Predictive.Analytics.business + Prescriptive.analytics.business +
    Big.Data.Business + Local.or.global.player + Internet.Activity.Score +
    average_L + average_M + struct_S + struct_U
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  15 -104.10
2  14 -106.06 -1 3.9184    0.04776 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood Ratio test with whole model: We used likelihood ratio test to verify hypothesis testing. Because the logit model is estimated using the maximum likelihood method. We verify the following hypothesis.

$$H0: B_1 = B_2 = B_3 \ldots = B_n = 0$$

The P value of the LR test is lower than 0.05. So we can reject the null hypothesis and all coefficients of logit modes are not equal to zero. They are significant.

Table: Result of likelihood Ratio test of logit model with all coefficients

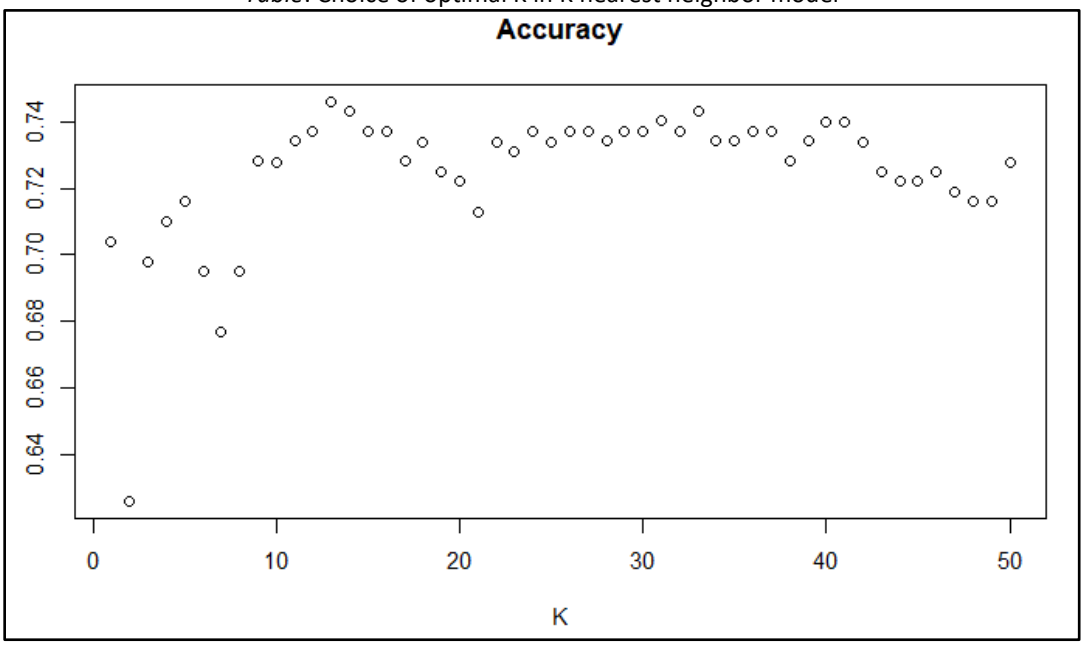
```
Likelihood ratio test

Model 1: Dependent.Company.Status ~ Have.been.part.of.successful.startups.in.the.past. +
    Consulting.experience. + Age.of.company.in.years + Machine.Learning.based.business +
    Predictive.Analytics.business + Prescriptive.analytics.business +
    Big.Data.Business + Local.or.global.player + Internet.Activity.Score +
    Age.of.company.in.years2 + average_L + average_M + struct_S +
    struct_U
Model 2: Dependent.Company.Status ~ 1
  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  15 -104.10
2   1 -215.01 -14 221.82  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.4 KNN model

In order to compare the performance of the logit model to other models, we used KNN model. We used 5 cross validation to find optimal hyperparameters. We assumed that k is from 1 to 50. We determined optimal k is 13 because when k is 13, accuracy is the highest.

Table: Choice of optimal K in K nearest neighbor model



6.5 Performance of Logit/probit and KNN

In this section, we compared the logit and KNN model's performance using a test set. Logit model's accuracy is higher than KNN model in test set. So we see that the Logit model predicts better.

Table: Accuracy of logit and K nearest neighbor models

	Train	Test
Logit (threshold=0.5)	0.87	0.69
KNN (k=13)	0.73	0.66

7. Conclusion and Findings

This study aims to assess the utility of two categories of startup success predictors, namely the characteristics of founders and the characteristics of the startup itself. The findings of this study deviate to a considerable extent from the existing research. Firstly, concerning the variables associated with the founders' characteristics category, it was unexpected to discover that both the level of education and its relevance to the venture had an insignificant impact on the startup. In comparison to previous research, our results exhibit partial alignment with the literature, as prior studies found a positive correlation between a high level of education and obtaining funds and capital, but found no significant impact on venture growth.

Regarding the founders' professional experience, it was intriguing to observe a positive relationship between founders with previous experience in medium-sized and large companies and startup success, while a negative relationship was observed between founders with prior startup experience and the success of the venture. This contrasting finding is noteworthy, as the literature review yielded two distinct and contradictory results regarding prior startup experience. In our study, we align more with Marino and Noble (1997), who did not find empirical evidence of significant influence of previous startup experience on the survival and growth of a startup.

Thus, in our primary hypothesis, only the founder's previous professional experience in medium and large companies is deemed significant, while the remaining variables are considered insignificant.

Regarding the variables associated with startup characteristics, it was observed that two selected variables based on existing literature, namely professional advisors and team size of employees, were found to be insignificant. Conversely, the type of products or services offered by startups, such as machine learning-based businesses, predictive analytic businesses, big data businesses, and those focusing on structured/unstructured data, were found to be significant and aligned with the existing literature. Current research supports the notion that businesses offering and delivering products/services in the growth stage positively impact startup success (Lussier et al., 2016).

Consequently, in our secondary hypothesis, it is concluded that only the type of business of the startup is significant, while other characteristics such as team size and the involvement of professional advisors are deemed insignificant.

8. Reference

Ries, E. (2011). The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses. Crown Business.

Picken, J. C. (2017). From startup to scalable enterprise: Laying the foundation. Business Horizons, 60(5), 587–595. http://dx.doi.org/10.1016/j.bushor.2017.05.002.

Axelson, U., Sorensen, M., & Strömberg, P. (2013). The alpha and beta of buyout deals. Working paper, Columbia University.

Lussier, R. N. (1995). A nonfinancial business success versus failure prediction model for young firms. Journal of Small Business Management, 33, 8-19.

Blank, S., & Dorf, B. (2012). The Startup Owner's Manual: The Step-by-Step Guide for Building a Great Company. BookBaby, Pennsauken.

Barney, J. B. (1986). Organization culture: Can it be a source of sustained competitive advantage? Academy of Management Review, 11, 656-665.

Mayer-Haung, K., Read, S., Brickmann, J., Dew, N., & Grichnik, D. (2013). Entrepreneurial talent and venture performance: A meta-analytic investigation of SME. Research Policy, 42, 1251-1273.

Zbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using crunchbase data. Information Processing and Management, 58, 102555. https://doi.org/10.1016/j.ipm.2021.102555

Carter, R., & Auken, H. (2006). Small firm bankruptcy. Journal of Small Business Management, 44, 493-512.

Colombo, M. G., & Grilli, L. (2005). Founders' human capital and the growth of new technology-based firms: A competence-based view. Research Policy, 34, 795–816. https://doi.org/10.1016/j.respol.2005.03.010

Hsu, D. H. (2007). Experienced entrepreneurial founders, organizational capital, and venture capital funding. Research Policy, 36, 722–741. https://doi.org/10.1016/j.respol.2007.02.022

Kalyanasundaram, G., Ramachandrula, S., & Hillemane, B. S. M. (2021). The life expectancy of tech start-ups in India: What attributes impact tech start-ups' failures? International Journal of Entrepreneurial Behavior Research, 27, 2050–2078. https://doi.org/10.1108/IJEBR-01-2021-0025

Spanjer, A., van Witteloostuijn, A., & van Witteloostuijn, B. A. (2017). The entrepreneur's experiential diversity and performance. Small Business Economics, 49, 141–161. https://about.jstor.org/terms

Cassar, G. (2014). Industry and startup experience on entrepreneur forecast performance in new firms. Journal of Business Venturing, 29, 137–151. https://doi.org/10.1016/j.jbusvent.2012.10.002

Marino, K. E., & Noble, A. F. D. (1997). Growth and early returns in technology-based manufacturing ventures. The Journal of High Technology Management Research, 8, 225–242. https://doi.org/10.1016/S1047-8310(97)90004-3

Brinckmann, J., & Högl, M. (2011). Effects of initial teamwork capability and initial relational capability on the development of new technology-based firms. Strategic Entrepreneurship Journal, 5, 37–57.

Lussier, R. N., & Maron, S. (2014). A Business Success versus Failure Prediction Model for Small Businesses in Israel. Business and Economic Research, 4, 63-81.

Lussier, R. N., & Hyder, S. (2016). Why businesses succeed or fail: a study on small businesses in Pakistan. Journal of Entrepreneurship in Emerging Economies, 8, 82-100.

Wojcik, P. (2023). KNN model.

9. Appendix

Appendix: R code

```
Sys.setenv(LANG = "en")
options(scipen = 5)

###-------------------------Packages-------------------------
library(tidyverse)
library(dplyr)
library(caret)
library("sandwich")
library("lmtest")
library("MASS")
library("mfx")
library("htmltools")
library("aod")
library(DescTools)
library(aods3)
library("logistf")

###-------------------------Data import-------------------------
library(readr)
CAX_Startup_Data_Dictionary <- read.csv2("CAX_Startup_Data_Dictionary.csv",header=T,sep=',') #
dictionary file of data set
df <- read.csv2("CAX_Startup_Data.csv",header = T,sep=',') # data set
df$Local.or.global.player=trimws(df$Local.or.global.player) # Removing white spaces from chosen
```

```
column
df=df %>% dplyr::select( # Target variable ### Initial variable selection by 4 categories
                  Dependent.Company.Status,
                  # Founder characteristics
                  Worked.in.top.companies,Average.size.of.companies.worked.for.in.the.past,
                  Have.been.part.of.startups.in.the.past.,
Have.been.part.of.successful.startups.in.the.past.,
                  Consulting.experience.,Was.he.or.she.partner.in.Big.5.consulting.,
                  # Startup characteristics
                  Age.of.company.in.years,Focus.on.private.or.public.data., Focus.on.consumer.data.,
                  Focus.on.structured.or.unstructured.data, Machine.Learning.based.business,
Predictive.Analytics.business,
                  Speech.analytics.business, Prescriptive.analytics.business,
                  Big.Data.Business, Product.or.service.company.,Local.or.global.player,
                  # Access to capital
                  Last.Funding.Amount,
                  # External factors
                  Number.of.Investors.in.Seed,Internet.Activity.Score,Number.of.Co.founders,
Number.of.of.advisors, Team.size.all.employees,
                  Number.of..Sales.Support.material,Number.of..of.Partners.of.company,
                 ) # selecting columns/variables

###-------------------------Data prepare-------------------------
# Categorical variables
df <- df %>% # Replace High value to high in chosen column
  mutate(Number.of..Sales.Support.material =
str_replace(Number.of..Sales.Support.material,"High","high"))
df <- df %>%  # Replace no value to Both value in chosen column
  mutate(Focus.on.structured.or.unstructured.data =
str_replace(Focus.on.structured.or.unstructured.data,"no","Both"))
df <- df %>%  # Replace not applicable value to Both value in chosen column
  mutate(Focus.on.structured.or.unstructured.data =
str_replace(Focus.on.structured.or.unstructured.data,"not applicable","Both"))
df <- df %>%  # Replace both applicable value to Both value in chosen columns
  mutate(Focus.on.structured.or.unstructured.data =
str_replace(Focus.on.structured.or.unstructured.data,"Botht applicable","Both"))
df$Local.or.global.player<-tolower(df$Local.or.global.player) # convert into lowercase or lower letter
of chosen column
for (i in c(2,3,4,5,6,7,9,10,11,12,13,14,15,16,17,18,19,26)) { # Replace No Info value of all categorical
variables by the most frequently value
  new_val<-names(which.max(table(df[,i])))
  df[,i][df[,i] == 'No Info'] <- new_val
}
for (i in c(2,3,4,5,6,7,9,10,11,12,13,14,15,16,17,18,19,26)) { # Replace no info value of all categorical
variables by the most frequently value
  new_val<-names(which.max(table(df[,i])))
  df[,i][df[,i] == 'no info'] <- new_val
```

```
}
df$Number.of..Sales.Support.material[df$Number.of..Sales.Support.material=='Nothing']<-
names(which.max(table(df$Number.of..Sales.Support.material))) # Replace Nothing value of chosen
column into the most frequently value
df$Number.of..Sales.Support.material=factor(df$Number.of..Sales.Support.material,ordered =
T,levels = c('low','medium','high')) # apply ordered factor type for chosen variable
df$Average.size.of.companies.worked.for.in.the.past=factor(df$Average.size.of.companies.worked.fo
r.in.the.past,ordered = T,levels = c('Small','Medium','Large')) # apply ordered factor type for chosen
variable
df$Number.of..of.Partners.of.company=factor(df$Number.of..of.Partners.of.company,ordered =
T,levels = c('None','Few','Many')) # apply ordered factor type for chosen variable
df=df %>% mutate_at(vars(Worked.in.top.companies,
                         Have.been.part.of.startups.in.the.past.,
                         Have.been.part.of.successful.startups.in.the.past.,
                         Consulting.experience.,
                         Product.or.service.company.,
                         Focus.on.private.or.public.data.,
                         Focus.on.consumer.data.,
                         Focus.on.structured.or.unstructured.data,
                         Local.or.global.player,
                         Machine.Learning.based.business,
                         Predictive.Analytics.business,
                         Speech.analytics.business,
                         Prescriptive.analytics.business,
                         Big.Data.Business,
                         Was.he.or.she.partner.in.Big.5.consulting.),as.factor) # apply not ordered factor type for
chosen variables
# Numerical variables
df=df %>% mutate_at(vars(Age.of.company.in.years,
                       Internet.Activity.Score,
                       Last.Funding.Amount,
                       Number.of.Investors.in.Seed,
                       Number.of.Co.founders,
                       Number.of.of.advisors,
                       Team.size.all.employees),as.numeric) # apply numeric type for chosen variables

# Missing values
colSums(is.na(df)) %>%
  sort() # sorted number of missing value of all columns
df$Average.size.of.companies.worked.for.in.the.past[is.na(df$Average.size.of.companies.worked.for.i
n.the.past)] <- names(which.max(table(df$Average.size.of.companies.worked.for.in.the.past)))
df$Number.of.Investors.in.Seed[is.na(df$Number.of.Investors.in.Seed)] <-
median(df$Number.of.Investors.in.Seed,na.rm = T) # fill in missing value of chosen variable with
median
df$Age.of.company.in.years[is.na(df$Age.of.company.in.years)] <-
median(df$Age.of.company.in.years,na.rm = T) # fill in missing value of chosen variable with median
df$Internet.Activity.Score[is.na(df$Internet.Activity.Score)] <-
```

```
median(df$Internet.Activity.Score,na.rm = T) # fill in missing value of chosen variable with median
df$Team.size.all.employees[is.na(df$Team.size.all.employees)] <-
median(df$Team.size.all.employees,na.rm = T) # fill in missing value of chosen variable with median
df$Last.Funding.Amount[is.na(df$Last.Funding.Amount)] <- median(df$Last.Funding.Amount,na.rm =
T) # fill in missing value of chosen variable with median

# Outliers
for (i in c(8,19,20,21,22,23,24)) { # All numeric variable's number of outlier using interquartile method
  q1=quantile(df[,i], .25)
  q3=quantile(df[,i], .75)
  IQR=IQR(df[,i])
  count_out<-subset(df, df[,i] > (q1 - 1.5*IQR) & df[,i] < (q3 + 1.5*IQR))
  print(paste0(colnames(df)[i]," variable count of outlier - ",count(df)-count(count_out)))
}
for (i in c(8,19,20,21,22,23,24)) {
  df[,i][df[,i] %in% boxplot(df[,i])$out] <- median(df[,i]) # Replace all outlier values by median
}


###------------------------Variable selection and create variable-------------------------
df$Dependent.Company.Status=ifelse(df$Dependent.Company.Status=='Success',1,0) # recoding
target variable
df$Dependent.Company.Status<-as.integer(df$Dependent.Company.Status) # recoding target variable
df$Team.size.all.employees2=df$Team.size.all.employees^2 # create new variable for checking non-
linear relationship
df$Age.of.company.in.years2=df$Age.of.company.in.years^2 # create new variable for checking non-
linear relationship
par(mfrow=c(4,2))
hist(df[,8],main = paste(colnames(df)[8])) # histogram of numeric variable
hist(df[,19],main = paste(colnames(df)[19])) # histogram of numeric variable
hist(df[,20],main = paste(colnames(df)[20])) # histogram of numeric variable
hist(df[,21],main = paste(colnames(df)[21])) # histogram of numeric variable
hist(df[,22],main = paste(colnames(df)[22])) # histogram of numeric variable
hist(df[,23],main = paste(colnames(df)[23])) # histogram of numeric variable
hist(df[,24],main = paste(colnames(df)[24])) # histogram of numeric variable
par(mfrow=c(1,1)) # Decided to remove following variables Number.of.Investors.in.Seed,
Number.of.of.advisors based on distribution of numerical variable
par(mfrow=c(4,5))
for (i in c(2,3,4,5,6,7,9,10,11,12,13,14,15,16,17,18,25,26,27)) { # Frequency distribution of categorical
variables
  barplot(prop.table(table(df[,i])),main = colnames(df)[i])
}
par(mfrow=c(1,1)) # Decided to Remove following variables based on frequency distribution -
Worked.in.top.companies, Was.he.or.she.partner.in.Big.5.consulting.,
"Speech.analytics.business","Number.of..of.Partners.of.company"
df=df %>% dplyr::select(-
c("Number.of.Investors.in.Seed","Number.of.of.advisors","Worked.in.top.companies","Was.he.or.she
```

```
.partner.in.Big.5.consulting.",
                "Speech.analytics.business","Number.of..of.Partners.of.company"))

###------------------------Data split into train and test-----------------------
df$Dependent.Company.Status<-as.factor(df$Dependent.Company.Status)
split <- createDataPartition(df$Dependent.Company.Status, p = 0.7,list = FALSE) # Divide data into
train 70%, test 30%
df_train <- df[c(split),] # train set
df_test <- df[-c(split),] # validation set
df=read_rds("df.rds") # We saved cleaned data as df.rds
split=read_rds("split.RDS") #
df_train=read_rds("df_train.rds") # we saved train set as df_train
df_test=read_rds('df_test.rds') # we saved test set as df_test
table(df_train$Dependent.Company.Status) # frequency of target variable in train set
table(df_test$Dependent.Company.Status) # frequency of target variable in test set

###-------------------------Data Recoding-------------------------
df_train$average_L=ifelse(df_train$Average.size.of.companies.worked.for.in.the.past=='Large',1,0) #
create new explanatory dummy variable with binary choice
df_train$average_M=ifelse(df_train$Average.size.of.companies.worked.for.in.the.past=='Medium',1,0
) # create new explanatory dummy variable with binary choice
df_train$average_S=ifelse(df_train$Average.size.of.companies.worked.for.in.the.past=='Small',1,0) #
create new explanatory dummy variable with binary choice
df_train$struct_S=ifelse(df_train$Focus.on.structured.or.unstructured.data=='Structured',1,0) #
create new explanatory dummy variable with binary choice
df_train$struct_U=ifelse(df_train$Focus.on.structured.or.unstructured.data=='Unstructured',1,0) #
create new explanatory dummy variable with binary choice
df_train$struct_B=ifelse(df_train$Focus.on.structured.or.unstructured.data=='Both',1,0) # create new
explanatory dummy variable with binary choice
df_train=df_train %>% dplyr::select(-
c(Average.size.of.companies.worked.for.in.the.past,Focus.on.structured.or.unstructured.data))

df_test$average_L=ifelse(df_test$Average.size.of.companies.worked.for.in.the.past=='Large',1,0) #
create new explanatory dummy variable with binary choice
df_test$average_M=ifelse(df_test$Average.size.of.companies.worked.for.in.the.past=='Medium',1,0)
# create new explanatory dummy variable with binary choice
df_test$average_S=ifelse(df_test$Average.size.of.companies.worked.for.in.the.past=='Small',1,0) #
create new explanatory dummy variable with binary choice
df_test$struct_S=ifelse(df_test$Focus.on.structured.or.unstructured.data=='Structured',1,0) # create
new explanatory dummy variable with binary choice
df_test$struct_U=ifelse(df_test$Focus.on.structured.or.unstructured.data=='Unstructured',1,0) #
create new explanatory dummy variable with binary choice
df_test$struct_B=ifelse(df_test$Focus.on.structured.or.unstructured.data=='Both',1,0) # create new
explanatory dummy variable with binary choice
df_test=df_test %>% dplyr::select(-
c(Average.size.of.companies.worked.for.in.the.past,Focus.on.structured.or.unstructured.data))
```

```
###-------------------------Descriptive analysis-------------------------
df_nums <-
  sapply(df, is.numeric) %>% # Correlation between numeric variables
  which() %>%
  names()
df_num_cor <-
  cor(df[, df_nums],
      use = "pairwise.complete.obs")


###-------------------------Model-------------------------
#--------------------Logit, Probit
logit <- glm(Dependent.Company.Status~.,data=df_train,family=binomial(link="logit")) %>% # logit
model with backward variable selection method
  stepAIC(trace = FALSE,direction = 'backward') # backward variable selection method
probit <- glm(Dependent.Company.Status~., data=df_train, # probit model with backward variable
selection method
          family=binomial(link="probit")) %>% stepAIC(trace = FALSE,direction = 'backward')  # backward
variable selection method
print(paste0(logit$aic,"  ",probit$aic)) # AIC - information criteria of logit and probit model

# logit, probit, lpm with the most important variables from backward variable selection method
logit <-
glm(Dependent.Company.Status~Have.been.part.of.successful.startups.in.the.past.+Consulting.experi
ence.+Age.of.company.in.years+

Machine.Learning.based.business+Predictive.Analytics.business+Prescriptive.analytics.business+Big.D
ata.Business+Local.or.global.player+

Internet.Activity.Score+Age.of.company.in.years2+average_L+average_M+struct_S+struct_U,data=df_
train,family=binomial(link="logit")) # logit model with most important explanatory variables
probit <-
glm(Dependent.Company.Status~Have.been.part.of.successful.startups.in.the.past.+Consulting.experi
ence.+Age.of.company.in.years+

Machine.Learning.based.business+Predictive.Analytics.business+Prescriptive.analytics.business+Big.D
ata.Business+Local.or.global.player+

Internet.Activity.Score+Age.of.company.in.years2+average_L+average_M+struct_S+struct_U,data=df_
train,family=binomial(link="probit")) # probit model with most important explanatory variables
lpm =
lm(as.numeric(Dependent.Company.Status)~Have.been.part.of.successful.startups.in.the.past.+Consu
lting.experience.+Age.of.company.in.years+

Machine.Learning.based.business+Predictive.Analytics.business+Prescriptive.analytics.business+Big.D
ata.Business+Local.or.global.player+
```

```
Internet.Activity.Score+Age.of.company.in.years2+average_L+average_M+struct_S+struct_U,data=df_
train) # linear probability model with most important explanatory variables
summary(logit) # summary of logit model
summary(probit) # summary of probit model
summary(lpm) # summary of lpm model

# Breausch-pagan test
bptest(logit) # breusch-pagan test of logit model
bptest(probit) # breusch-pagan test of probit model
bptest(lpm) # breusch-pagan test of lpm model

# Marginal effects with average characteristics
(meff =
logitmfx(Dependent.Company.Status~Have.been.part.of.successful.startups.in.the.past.+Consulting.e
xperience.+Age.of.company.in.years+

Machine.Learning.based.business+Predictive.Analytics.business+Prescriptive.analytics.business+Big.D
ata.Business+Local.or.global.player+

Internet.Activity.Score+Age.of.company.in.years2+average_L+average_M+struct_S+struct_U, data =
df_train, atmean=TRUE))

# R squared of logit model
PseudoR2(logit,which = c('all')) # all R squared
countR2<-function(m) mean(m$y==round(m$fitted.values))
countR2(logit) # count R squared of logit model
adj.countR2<-function(m) {
  n<-length(m$y)
  k<-max(table(m$y))
  correct<-table(m$y==round(m$fitted.values))[["TRUE"]]
  (correct-k)/(n-k)
}
adj.countR2(logit) # Adjusted count R squared of logit model

# LINKTEST
source("linktest.R")
linktest_result = linktest(logit)
summary(linktest_result)

# Likelihood Ratio test with 1 variables
logit_rest <-
glm(Dependent.Company.Status~Have.been.part.of.successful.startups.in.the.past.+Consulting.experi
ence.+Age.of.company.in.years+

Machine.Learning.based.business+Predictive.Analytics.business+Prescriptive.analytics.business+Big.D
ata.Business+Local.or.global.player+
```

```
Internet.Activity.Score+average_L+average_M+struct_S+struct_U,data=df_train,family=binomial(link="logit"))
lrtest(logit, logit_rest)

# Likelihood Ratio test with all variables whole model
null_logit<-glm(Dependent.Company.Status~1,data=df_train,family=binomial(link="logit"))
lrtest(logit, null_logit)



#--------------------KNN
ctrl_cv5 <- trainControl(method = "cv",number = 5) # cross validation
grid <- expand.grid(.k=seq(1,50,by=1)) # search space of hyperparameters
df_trainknn=df_train %>% dplyr::select(-
c("Have.been.part.of.startups.in.the.past.","Focus.on.private.or.public.data.","Focus.on.consumer.data.",

"Product.or.service.company.","Last.Funding.Amount","Number.of.Co.founders","Number.of..Sales.Support.material")) # data prepare for knn model
knn_train <- train(Dependent.Company.Status ~ ., data = df_trainknn,method = "knn", trControl = ctrl_cv5, tuneGrid = grid) # hyperparameter tuning
plot(knn_train$results$k, knn_train$results$Accuracy, ylab="",xlab="K", main="Accuracy") # optimal k
k<-knn_train$results$k[knn_train$results$Accuracy==max(knn_train$results$Accuracy)] # optimal k
k <- knn_train$bestTune #  optimal k
knn_model <- train(Dependent.Company.Status ~ ., data = df_trainknn, # train knn model with optimal k
      method = "knn", trControl = ctrl_cv5, tuneGrid = k)

#--------------------Performance metrics of logit and KNN model
# Predicted value
predict_logit_train <- logit %>% predict(df_train, type = "response") # predicted value of logit on train set
predict_logit_train <- as.factor(ifelse(predict_logit_train  > 0.5, 1, 0))
predict_logit_test <- logit %>% predict(df_test, type = "response") # predicted value of logit on test set
predict_logit_test <- as.factor(ifelse(predict_logit_test  > 0.5, 1, 0))
predict_knn_train<-predict(knn_model,df_train) # predicted value of KNN model on train set
predict_knn_test<-predict(knn_model,df_test) # predicted value of KNN model on test set
# Confusion matrix
confusionMatrix(predict_logit_train, df_train$Dependent.Company.Status) # performance matrix of predicted value of logit on train set
confusionMatrix(predict_logit_test, df_test$Dependent.Company.Status)# performance matrix of predicted value of logit on test set
confusionMatrix(predict_knn_train, df_train$Dependent.Company.Status)# performance matrix of predicted value of KNN on train set
confusionMatrix(predict_knn_test, df_test$Dependent.Company.Status)# performance matrix of predicted value of KNN on test set
```