National College of Ireland

Indian Start-ups' Success Prediction Using Machine Learning

MSc Research Project
Data Analytics

Nixon Balu
Student ID: x20247788

School of Computing
National College of Ireland

Supervisor: Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Nixon Balu		
Student ID:	x20247788		
Programme:	M.Sc. Data Analytics	**Year:**	2022
Module:	Research Project		
Supervisor:	Christian Horn		
Submission Due Date:	1st February 2023		
Project Title:	Indian Start-up's Success Prediction Using Machine Learning		
Word Count:	4865	**Page Count:**	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Nixon Balu

Date: 31st January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Indian Start-ups' Success Prediction Using Machine Learning

Nixon Balu
x20247788

Abstract

Entrepreneurship is essential to the growth of the economy. Their success generates international investment and expands job opportunities. Investors stand to gain significantly from the success of these high-risk businesses, which in turn enables them to support further start-ups and maintain the economic growth cycle. In order to create a trustworthy and impartial model to predict the performance of such start-ups in the Indian market, this research introduces a novel approach in the feature engineering space. Furthermore, two experiments with different sampling techniques are designed, one of which uses data from start-ups only in India and the other of which includes start-ups from all over the world. These models are assessed with the state-of-the-art evaluation metrics used in this domain. Random Forest with weighted balance sampling technique demonstrated to be the most efficient model having significant and consistent F1, recall and accuracy scores.

1 Introduction

The number of start-ups fostered in India has reached over 50,000, and it is constantly increasing by 15% yearly (Chaudhari & Sinha 2021). As of June 2021, start-ups in India had raised more than $10 billion, placing it third in the world in the start-up ecosystem (Choudhary et al. 2022). The primary objective of many investors, particularly the larger Venture Capital (VC) and Private Equity (PE) firms, is to invest in promising early-stage start-ups. They work by utilizing their resources, including their assets, market knowledge, and skills, to acquire a stake in start-up businesses with the hope of reaping significant future financial rewards. Early-stage startups tend to be more unpredictable because there is less previous data to base assessments on, which makes them extremely subjective. In order to

choose the best start-ups, investors must carefully consider the risks and potential returns of each investment. This takes a lot of time and labor.

This study is driven by the fact that large sums of money are brought in by investors, particularly the VC and PE firms, through foreign direct investments (FDI). FDI plays a significant role in determining a nation's rate of economic expansion. Therefore, it is essential to have a strong startup evaluation mechanism so that India as a nation can benefit from these investments.

Numerous studies have been done in the past to examine the factors that contribute to a business' success. While some of them utilize more conventional human surveys, others use more advanced machine learning approaches. The main contribution of this research is the novel approach in the feature engineering space that reduces bias introduced in making predictions.

The goal of this study is to evaluate the success of start-ups using supervised machine learning techniques and, more importantly, to answer the Research Question: "To what extent can supervised machine learning models be used to predict the success or failure of Indian start-up's?". Additionally, the emphasis is placed on the features used to make these predictions.

The rest of the paper follows the following format: A summary of the Related Work is discussed in Section 2, the Methodology is emphasized in Section 3, the Project Design and Implementation is discussed in Section 4, the Results and Discussion is addressed in Section 5, and the Conclusion and Future Work is covered in Section 6.

2 Related Work

Numerous researchers are interested in investors because of their considerable impact on the launch of new companies. The top peer-reviewed articles in this field from the previous five years were assessed, and evidence and research gaps were identified.

2.1 Ensembles Models and Frameworks in Predicting Success of Start-ups

To help investors choose which early-stage companies to invest in, the ESI framework offers a checklist (Corea et al. 2021). They rank the organizations using a multivariate discriminant analysis and, in the absence of financial data, they make suggestions based on qualitative data. If the business received further funding and had the potential for an exit through an acquisition or IPO, those factors were taken into account as a measure of success. They chose businesses that were under four years old and had not yet received series C funding. The study used a gradient tree boosting classifier and produced a list of 21 features, including traits of the founders, traits centered on employment, and traits centered on investment. The checklist, however, misses aspects like market size and external factors that could be quite crucial from the perspective of an investor. Similar to this, the CaptalVX framework proposed by Ross et al. (2021), which combines Deep Learning, Random Forest, XGBoost, and KNN models, determines if a startup's outcome was successful based on the possibility of an exit or whether it remained private. They present a two-stage classification process, measuring the startup's exit scenario in the first step and determining whether there was room for more fundraising rounds in the second. Although this system outperformed the individual models in terms of performance, its deep learning component gave it a "black box" quality. Because of this, it is challenging to evaluate the model's performance on particular attributes.

Another interesting study involves conducting web searches on the names of startups and combining the results with the structured data already available to determine the success of such startups (Sharchilev et al. 2018). It does this by carefully examining the data of only those businesses that have already secured their first round of investment and determining whether they make progress toward securing subsequent rounds within a predetermined time frame. Greater investment rounds indicate that the business is more established, hence this provides a practical measurement of success. The created framework incorporates the features from the final CatBoost model into the predictions from Logistic Regression and Neural Networks. Lin (2019) introduces a platform for artificial intelligence that enables startups to forecast market response before signing up on a Taiwanese equity crowdfunding platform called GISA. To be eligible for registration in GISA, start-ups must obtain a 100% equity subscription rate. It displays how the startup's concept or business strategy has been received by the market. Access to essential

financial data was possible because the study is narrowly focused on a particular market. For the larger startup ecosystem, such information is not accessible to the broader public.

Sherk et al. (2019) investigate a computer approach for forecasting whether the start-ups on the SharkTank show would strike a deal with the sharks. The computational model investigates a linear prediction model regarding whether the trade will occur or not before making an effort to reduce the error between the prediction and the actual data. Later on, l2 regularization is carried out to prevent overfitting, which ultimately leads to a non-negative square problem that is resolved using the MATLAB Isqonneg toolbox. Compared to other adapted models in this field, the model produced better outcomes. Instead of the considerably larger VC or PE organizations, this structure only serves the start-up market, which works with direct individual investors.

2.2 Data Choice and Feature Selection in Start-up Ecosystem

The study by Zhang et al. (2017) clarifies that a startup's survival and subsequent success depend on its ability to raise money. Their data shows a high correlation between successful crowdfunding and active social media participation. They used APIs to gather information over a period of 7 to 10 months from different social media platforms, including Facebook, Twitter, and AngelList. As a result of sampling approaches being used to address the imbalanced data that had been collected, accurate predictions were obtained. Interestingly, Lee et al. (2018) solely examined the Technology category after compiling 216,136 pages of Kickstarter.com crowdfunding campaigns between April 2009 and August 2017. Only the features that were focused on project idea, project progress, and comments that were included in both were extracted because the data was in picture and video format. Additionally, speech information was taken from political campaign films. The suggested model was able to predict successful fundraising with state-of-the-art accuracy after looking at the text data from these features.

In Saura et al. (2021), the data is extracted from Twitter chat in three steps. First, a sentiment analysis is conducted on the data and the sentiments are categorized into groups. Next, a Latent Dirichlet Allocation model is created, which divides the tweets into topics using user-generated content from particular tweets like #IndianStartups and similar hashtags. The key features of these topics are then determined using text analysis. This study does not include any information on

actual investments; it only focuses on descriptive statistics for investors. Snehal et al. (2020) uses a different method of data selection and chooses 1286 new startups that were supported by the top two accelerator programs in India, the USA, and Brazil. Four important criteria—growth, survivability, acquisition, and financial characteristics—were used by the authors to evaluate the success of these businesses. The variations in the regional ecosystems had a big impact on the outcomes. The study by Antretter et al. (2019) shows how online legitimacy, a metric of social acceptance based on Twitter content, can be used to accurately predict whether new enterprises will survive. In order to reduce sampling biases during the data collection phase, they use stratified bootstrapping. The article by Li (2020) uses 40 direct features from Crunchbase data and compares two machine learning methods, namely random forest and support vector machines.

Reviewing the various methods for data and feature selection reveals that the vast majority of studies ignored funding details and geographic characteristics. Investors place a high value on these traits because they play a crucial role in identifying successful start-ups.

2.3 Role of Biases in Start-up Ecosystem

In Żbikowski & Antosiuk (2021), the authors discuss on how to evaluate a start-up's performance without bias. As the final list of features, they only use data that was accessible at the company's founding. Web crawling is another technique they use to evaluate the firm homepage URLs and determine whether they are accessible. The company's operational nature is indicated by an active homepage URL. For the purpose of minimizing bias, the authors define a start-up as successful if it has obtained at least series B funding, has been acquired, or has successfully exited through an initial public offering (IPO). Despite having a high accuracy and precision score, this approach had considerably low recall and F1 scores. This shows that one of the classes may have been incorrectly classified. Blohm et al. (2022) makes an interesting attempt to contrast investor investment returns using data from business angels (BA) and machine learning (ML) algorithms. They assess the BAs' investment platform rigorously and train the ML algorithm with the same data to eliminate knowledge gaps. The idea that the ML algorithm outperformed the BA's based on these criteria was also critically examined with regard to the decision biases of the BAs, such as overconfidence, local bias, and loss aversion. The only circumstance in which the BAs outperformed the ML algorithm was when they had a wealth of investment experience.

Another study by Taboga (2022) talks about how start-up VC financing rounds are biased in terms of size. It was noted that investment amounts provided to start-ups by VCs varied between nations. This resulted from the striking variations in each funding cycle. According to their data, significant sums of money were injected into promising start-ups in nations where venture capitalists were the dominant source of funding, making it extremely impossible for other start-ups to receive any funding at all. Because of this, there were prejudices present during national VC fundraising rounds, which affected how a start-up developed. Arroyo et al. (2019) used a time-aware methodology to divide the dataset into warm-up and simulation stages to address time-sensitive biases. Only businesses that had received Series B or lower capital or those that had received no funding before the simulation window opened were selected. It's probable that many successful enterprises were excluded from the training set as a result of the average 1.5-year gap between venture rounds B and C.

3 Methodology for Success Prediction of Indian Start-ups

The methodology used for this research is based on the Cross Industry Standard Process for Data Mining (CRISP-DM) framework without the deployment phase. Fig. 1 illustrates the various steps that are included and will be further explained in the next section.

3.1 Business Understanding

Before embarking on any analytical task, it's crucial to understand what will be accomplished with it. To make the most use of the resources at hand, it is common practice in the fields of data analytics and data mining to have a business objective. In this research, an effort is made to comprehend the importance of start-up's success in order to ensure profitable investment opportunities for investors.

3.2 Data Understanding

The data for this research was obtained from Crunchbase, which is one of the largest data collection points for private and public companies, which holds information on the founders, the funding rounds, investors, among others. The information is broken down into various CSV files, including ones for organizations,

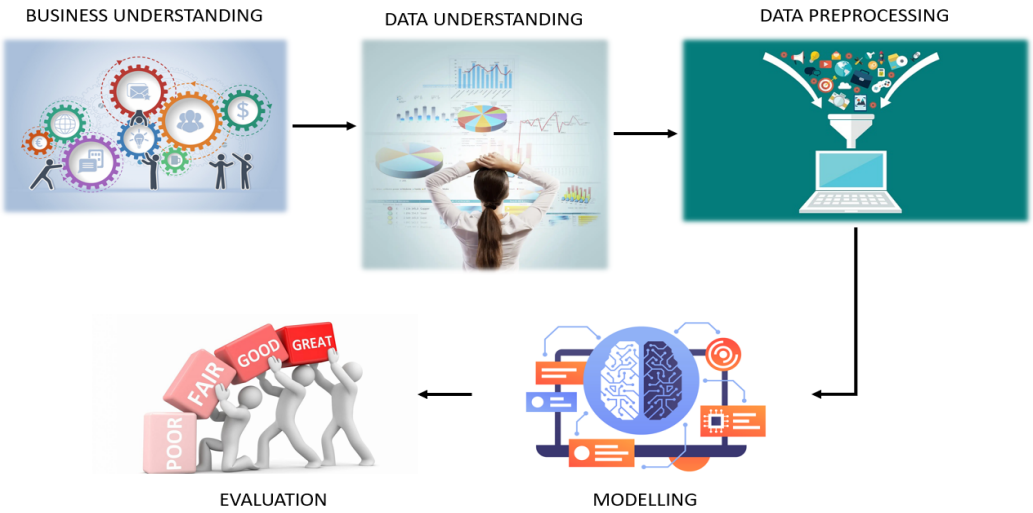

Figure 1: Methodology for Success Prediction of Indian Start-ups

fundraising rounds, ipos, and acquisitions, each of which contains information under the name given to it. For instance, the CSV file organizations contains information about specific businesses, like their legal name, country code, category they fall under, and staff count, to mention a few.

Only two of these CSV files—those pertaining to organizations and financing rounds—will be used in this study. The files are individually loaded in separate Pandas dataframes and only few attributes from each dataframe are used, thus creating a single final dataframe. This is accomplished by utilizing Pandas' MERGE function, which promotes the idea of joins. Figure 1 depicts the connection made between each dataframe.

3.3 Data Preparation

Initially, errors, discrepancies, and missing values are removed from the data. Clean data promotes the highest level of decision-making quality and improves model performance. Data merging will be done when it has been cleaned, and feature engineering will follow thereafter. The data will then be scaled and processed to improve the comparability of the data points.

3.4 Modelling

Three of the most prominent ML algorithms in this domain which are Random Forest, Decision Trees and Logistic Regression, will be evaluated using these balancing techniques.

Logistic Regression

One of the most used classification algorithms is logistic regression. The Maximum Likelihood estimation approach is used in place of the Ordinary Least Squares estimation approach in this case. In order to increase the likelihood of achieving the observed results, this estimation employs an iterative approach to assess the magnitude and direction of the coefficients.

Decision Trees

Decision trees are supervised machine learning techniques where training data is continuously divided based on a specific parameter. The tree can be explained using two elements: decision nodes and leaves. The leaves represent the outcomes. The decision nodes are where the final classification is done, and this is where the data is separated (Charbuty & Abdulazeez 2021).

Random Forest

An ensemble method made up of several individual decision trees is called random forest. Each tree makes a forecast for a certain class, and the prediction made by the model is based on the class receiving the most votes.

3.5 Evaluation

The majority of the work done in the subject of ML explanation research is focused on creating novel tactics and approaches that will improve predictability while aiming to stop the loss in prediction accuracy. Although different theories have been considered, it is still unclear which theory is best suited for a particular ML solution in a particular context and for a certain domain expert (Zhou et al. 2021).

Only measures that have been widely utilized in this field are included in the

evaluation phase of this study. These metrics include F1 score, recall, precision and accuracy.

Recall

The model recall score assesses the model's ability to accurately forecast positives out of actual positives. In this study, recall score is calculated using the following formula:

$$Recall = TP_S/(FN_S + TP_S) \tag{1}$$

where,
TP_S represents the start-ups that are correctly classified as successful, and
FN_S represents successful start-ups that have been classified as failures

Precision

The precision score assesses how precisely positive cases are predicted. The calculation for the precision score is shown below.

$$Precision = TP_S/(FP_S + TP_S) \tag{2}$$

where,
FP_S represents failed start-ups that are incorrectly classified as successful.

F1 Score

The recall and precision scores are used to calculate the F1 score, although both of them are given equal weights. It is frequently used in place of accuracy. The formula for F1 score is

$$F1Score = 2 * Recall * Precision/(Recall + Precision) \tag{3}$$

4 Project Design and Implementation

The project is implemented using Jupyter Notebook. The data for the research is legally obtained from Crunchbase. A thorough selection of the necessary CSV files was made from among the different CSV files that the dataset contained, and these files were then subject to pre-processing. New features are added to the

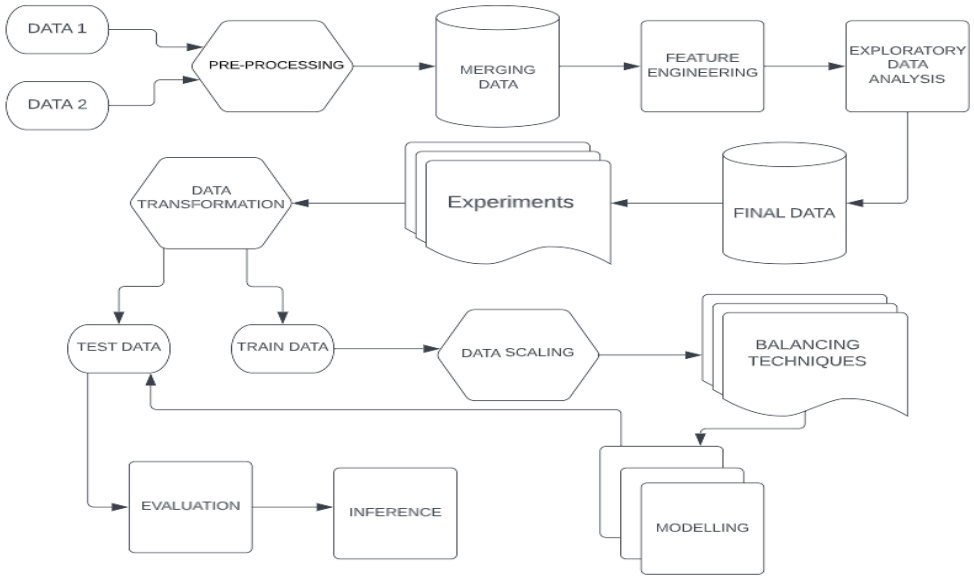

Figure 2: Project Design

processed data, which is then examined for intriguing patterns. Additionally, the final data was transformed as needed before being fed into the different experiments under study. The project design is illustrated in Fig. 2 and will be further discussed below.

4.1 Data Pre-processing

Although the data owner updates the material frequently, there are some inconsistencies in the data that is currently available. This is overcome by means of meaningful data pre-processing techniques. First, only the entries from the attributes – investment_type, announced_on, raised_amount_usd, and org_name are picked from the funding rounds file, which is then loaded into a pandas dataframe. Only four of the many sorts of investments for organizations listed in the variable investment_type were selected since they are generally regarded as prospective investments that can help decide a start-up's success. From each of the four investment categories—seed, series A, series B, and series C—the date of investment, the type, and the amount raised are retrieved separately. Duplicate entries are dropped and all these dataframes are sorted by the investment date to arrange the organizations in the order of oldest to the newest.

To prevent any information loss, each of the four dataframes is merged with the others in a logical sequence. The companies that have received their seed and series A funding rounds are the first to merge. Due to the fact that some companies have secured series A funding without receiving seed money, this merge is carried out using an outer join. The next step is to combine this fresh dataframe with the companies that have already secured series B funding. Since it makes sense to believe that businesses advance to series B funding only after securing the series A funding, left join is utilized in this merger. Similar steps are taken when combining this dataframe with companies that have obtained series C funding.

Next, the file for organizations is selected, and only the attributes - name, country_code, founded_on, category_list, total_funding_usd, and status are chosen to generate a new dataframe. As it includes a complete list of companies and information on their status, the organizations dataframe is regarded as the primary dataframe. Duplicate entries are eliminated, same as previously.

By combining the organizations dataframe with the previously merged funding rounds dataframe using a left join, a new dataframe is created. The records of companies in this dataframe are replaced with a 1 if they have received funding of any kind, or if they have exited through an acquisition or an IPO, and a 0 if not. Only the companies from the data that have no missing values for the total investment received are picked in order to achieve a balance between reducing loss of valuable information and preserving good quantity of data for effective machine learning. The missing value treatment is now applied to the data, and some are simply eliminated while others are either replaced with the corresponding mean value or 0.

4.2 Feature Engineering

By using the attributes from the resulting clean dataframe, new features are produced using feature engineering. The first feature is introduced by taking the individual categories from the category_list column using the str.split()method and assigning them to a new variable called category.

The next step is the conversion of all attributes that include dates to datetime datatypes. This conversion enables the creation of new attributes, such as founding_to_seed_months, founding_to_seriesA_months, seriesA_to_seriesB_months, and

Table 1: Description of Attributes before Feature Engineering

Feature	Description	Datatype
country_code	country of founding	object
founded_on	date of founding	object
category	industry category	object
total_funding_usd	total funding raised in USD	float
seed	if seed funding received	object
seed_date	date of seed funding	object
raised_amount_seed	total seed amount raised in USD	float
seriesA	if series A funding received	object
raised_amount_seriesA	total series A amount raised in USD	float
series_a_date	date of series A funding	object
seriesB	if series B funding received	object
raised_amount_seriesB	total series B amount raised in USD	float
series_b_date	date of series B funding	object
seriecC	if series C funding received	object
raised_amount_seriesC	total series C amount raised in USD	float
series_c_date	date of series C funding	object
status	is company operating, acquired, ipo or closec	object

seriesB_to_seriesC_months, each of which reflect the relevant period measured in months starting from the date of founding. Few records in these new features have negative or extremely high values (greater than 10000). The negative values have no significance because no companies can secure any kind of financing before it is even created. Therefore, these records are dropped. Other instances involve values that are far too high to be considered realistic. This is addressed by grouping them into sensible ranges of 1–12, 12–24, 24-36, and 36 months and above. There is a separate category for values that are zero, indicating that these companies have not received any funding. With this approach, the previously developed features are replaced with this categorical set of new features, which are seed_duration, to_seriesA, seriesA_to_B, and seriesB_to_C.

Additionally, a new feature was created to assesses a company's success based on two criteria. The first criteria dictates that the company needs to be operational and have at least series B level funding. A series B funding suggests that the company has grown its user base significantly, surpassed the seed and series A investment goals, and is now positioned to be valued between $30 million and $60 million. The second criteria is the company's ability to exit through an acquisition or an IPO, which is another metric for success. All other instances were viewed as failures. The attributes are then renamed for easier comprehension. The final dataframe has 220500 records and 16 attributes. The description of attributes

Table 2: Description of Attributes after Feature Engineering

Feature	Description	Datatype
country_code	country of founding	object
category	industry category	object
seed	if seed funding received	category
total_funding_usd	total funding raised in USD	float
seed_duration	category of seed duration	category
seriesA	if series A funding received	category
raised_amount_seed	total seed amount raised in USD	float
to_seriesA	category of founding to series A	category
raised_amount_seriesA	total series A amount raised in USD	float
seriesB	if series B funding received	category
seriesA_to_B	category of series A to series B	category
raised_amount_seriesB	total series B amount raised in USD	float
seriesC	if series C funding received	category
seriesB_to_C	category of series B to series C	category
raised_amount_seriesC	total series C amount raised in USD	float
is_successfull	whether company is successful or not	object

before and after feature engineering is shown in table 1 and table 2 respectively.

4.3 Data Visualization

The data was studied to identify interesting patterns. Fig. 3 corresponds to the top 10 countries in the world where most start-ups originate from. With about 45% of the total figure, the United States of America comes in first. Besides the United States, other nations with a large number of start-ups include China, the United Kingdom, Canada, and India.

The top 10 industries for start-ups worldwide and in India are depicted in Fig. 4. Biotechnology, healthcare, e-commerce, and artificial intelligence are among the top 10 industries that the majority of start-ups target. It's interesting to note that when this is compared to Indian start-ups, there are a few additions and exclusions. India does not have the same level of popularity in the biotechnology sector as the rest of the world. India also has a larger presence in the food and beverage and agriculture sectors than the rest of the world.

Fig. 5 displays the current status of start-ups around the world and in India. It demonstrates that roughly 73% of start-ups are still in business, 13% have shut down, and the other 13% have been acquired or have made their exit through an

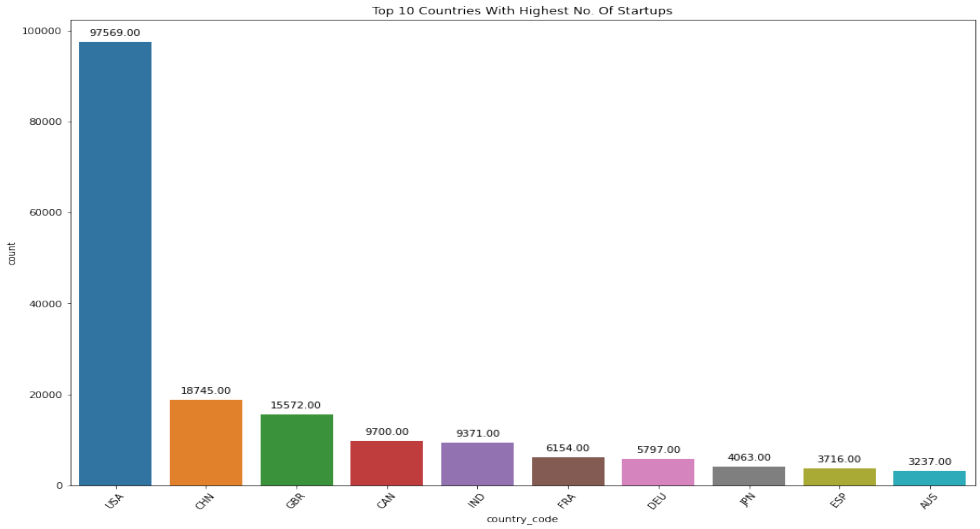

Figure 3: Top 10 Countries with Highest No. of Start-ups

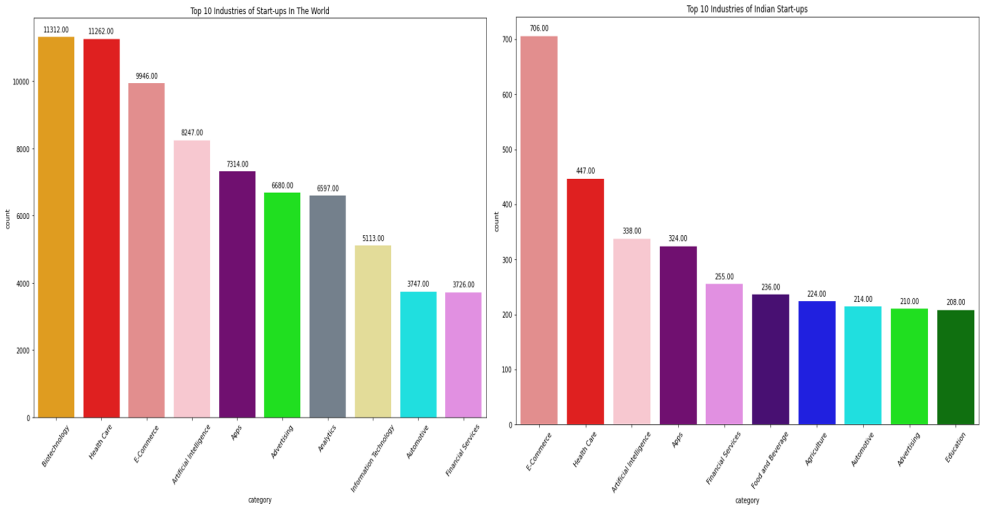

Figure 4: Top 10 Industries of Global vs. Indian Start-ups

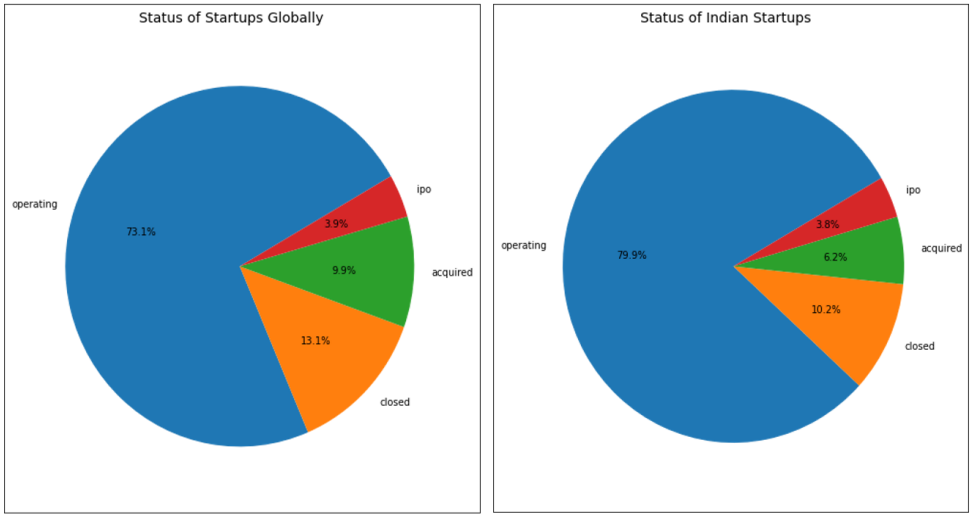

Figure 5: Current Status of Global vs. Indian Start-ups

IPO. It is interesting to note that only about 10% of Indian start-ups have closed their doors while approximately 80% are still in business. The proportion of Indian businesses that are bought out or are issued into an IPO, however, is only 10%.

4.4 Data Transformation

The data is transformed to appropriate datatypes. In order to efficiently support machine learning, attributes that belong to the object and category datatypes are label encoded and transformed into numerical data. Two dataframes are created from the transformed data, one containing just the target variable and the other including all other attributes. Following this, the test size is set to 20% for the train-test split using the train_test_split() method.

To prevent data leakage, scaling is carried out with the split data using the StandardScalar() method, in which the train dataset is fitted with the scalar which is then used to transform the test data.

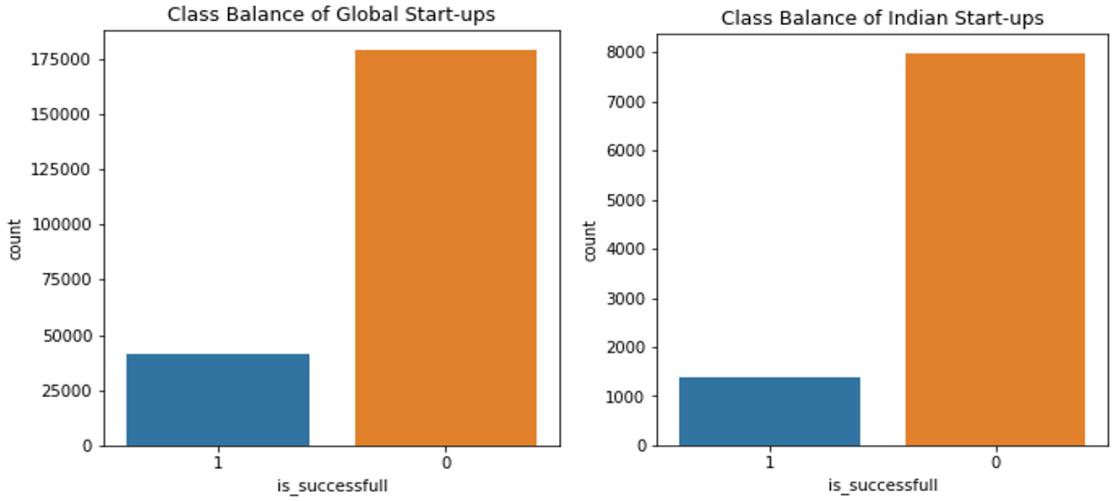

Figure 6: Class Balance of Global vs. Indian Start-ups

4.5 Model Building

It is clear from Fig. 6 that the target variable's distribution between the two classes is extremely skewed. This is a common scenario in the start-up industry as most companies fail within the first year of their founding. The modelling phase produces subpar results when employing this imbalanced data because machine learning algorithms struggle to handle them. As a result, the model building process is introduced with balancing techniques to address it.

Two experiments are conducted to determine differences in factors that influence a start-up's success. Different balancing techniques are used in each of these experiments. Three of the most prominent ML algorithms in this domain which are Random Forest, Decision Trees and Logistic Regression, will be evaluated using these balancing techniques.

Experiment 1: Success Prediction of Start-ups from Around the World

The scaled and transformed data which consists of start-ups from around the world is used for this experiment. Three balancing techniques are employed to correct the imbalance in the target variable.

Adjusting Class Weights to Balanced:

This method entails balancing the classes using the class_weight parameter in each machine learning model under study. This parameter's value is set to "balanced," which adjusts the class weights for both classes during the training process. The formula below serves as the basis for calculating these weights.

$$weight_c = t\_samples/(t\_classes * t\_samples_c) \tag{4}$$

where,
$weight_c$ is the weight of the class
$t\_samples$ is the total no. of records in the data
$t\_classes$ is the no. of unique classes
$t\_samples_c$ is the no. of records in the respective class

Systematic Minority Oversampling Technique (SMOTE):

Oversampling methods like SMOTE are preferred if the minority class has incredibly few records. The SMOTE algorithm adds new members to the minority class to balance out the majority class. In order to accomplish this, synthetic data points are created using the knowledge gained from the original data points. SMOTE selects examples that are close to one another in the feature space, draws a line connecting the examples, and then creates a new sample at a point along the line. In particular, a random representative of the minority class is first picked. Then, for that example, k nearest neighbours are located. A synthetic instance is created at a randomly chosen point in feature space between the two examples, using a neighbour that was chosen at random as its neighbour.

Random Undersampling:

This technique seeks to balance the imbalance between the two classes by choosing random samples from the majority class and disregarding them. When there is a sizable amount of data in the minority class that supports effective training of the data, this strategy appears to be more effective.

Experiment 2: Success Prediction of Indian Start-ups

The final data obtained after feature engineering is taken into account, and only Indian start-ups are selected. The country_code attribute is removed from the

dataframe. Label encoding is done on the object and category datatypes to transform them into numerical data. The transformed data is split into two dataframes, one with just the target variable and the other with all attributes. The test size is then adjusted for the train-test split to be 20%. Similar to before, scaling will be performed on the train dataset to prevent data leaking.

The large class disparity accounting to low number of successful start-ups is still clearly visible. Therefore, only two balancing techniques are used in this experiment because of the availability of fewer records in the minority class. The techniques used in this experiment includes SMOTE and adjusting class weights to balanced.

The results of models evaluated using both these experiments are tabulated and discussed in the next section.

5 Results and Discussion

This research places a strong emphasis on the value of feature selection in creating effective machine learning models for determining start-up's success. For start-ups around the world, three ML algorithms and three balancing techniques have been employed, and the same algorithms are used for Indian start-ups but with only two balancing techniques.

To determine the best model, different evaluation metrics were applied. Table 3 and Table 4 displays the results of all the models that were created. It is clear from both experiments that of all the balancing techniques used, Logistic Regression produces the best F1, accuracy, and precision scores. However, recall scores are substantially lower.

Even though F1 and accuracy scores are better, it is crucial to correctly identify successful start-ups since misclassifying failing start-ups as successful, causes significant losses for investors. This inaccurate evaluation concerns the investors, who then have a tendency to sell their positions in that investment as quickly as possible in order to limit their losses. Employees are negatively affected by this, and economies as a whole are adversely affected.

A model must therefore have good F1 and accuracy scores as well as good recall

Table 3: Results of ML Models for Success Prediction of Global Start-ups

Model Name	F1 Score	Accuracy	Recall	Precision
LR with Wighted Balance	0.836	0.844	0.469	0.604
LR with SMOTE	0.843	0.843	0.464	0.647
LR with Random Under Sampling	0.841	0.851	0.454	0.642
RF with Wighted Balance	**0.834**	**0.837**	**0.511**	**0.571**
RF with SMOTE	0.813	0.804	0.608	0.479
RF with Random Under Sampling	0.781	0.762	0.673	0.414
DT with Wighted Balance	0.789	0.781	0.517	0.428
DT with SMOTE	0.775	0.759	0.587	0.401
DT with Random Under Sampling	0.726	0.695	0.666	0.338

Table 4: Results of ML Models for Success Prediction of Indian Start-ups

Model Name	F1 Score	Accuracy	Recall	Precision
LR with Wighted Balance	0.879	0.899	0.341	0.912
LR with SMOTE	0.884	0.892	0.476	0.688
RF with Wighted Balance	**0.866**	**0.872**	**0.465**	**0.575**
RF with SMOTE	0.838	0.829	0.561	0.433
DT with Wighted Balance	0.854	0.859	0.439	0.517
DT with SMOTE	0.817	0.803	0.561	0.379

scores. When compared to all the models in both experiments, it can be concluded that Random Forest with weighted balancing technique has a good balance of all these scores, which is consistent with a reliable machine learning model. It has recall scores for global and Indian start-ups at 51% and 46.5% respectively.

6 Conclusion and Future Work

This study addresses the lack of a reliable and bias-free methodology to assess the success prediction of Indian start-ups by carefully building on previous studies. An attempt is made to compare Indian and international start-ups in order to ascertain the differences that geography and funding variables will have on the measure of success on these start-ups.

This study aims to attract investors to investment opportunities by highlighting the key elements of the Indian startup ecosystem. It is challenging for investors to decide which start-ups to invest in since the start-up ecosystem is so unpredictable given the variety of risks involved. The goal of the research is to address this problem and provide investors with the best bets for making wise and strategic investment decisions.

The final processed data used in this research was very huge and in the tune of 1.5 GB. The use of text features such as company description and founder's characteristics was hampered by a lack of time and technical restrictions. In the future, text features will also be included.

Acknowledgement

I would like to express my deepest gratitude to Prof. Christian Horn for his constant support and guidance with the submission of this research.

References

Antretter, T., Blohm, I., Grichnik, D. & Wincent, J. (2019), 'Predicting new venture survival: A twitter-based machine learning approach to measuring online legitimacy', *Journal of Business Venturing Insights* **11**, e00109.

Arroyo, J., Corea, F., Jimenez-Diaz, G. & Recio-Garcia, J. A. (2019), 'Assessment of machine learning performance for decision support in venture capital investments', *IEEE Access* **7**, 124233–124243.

Blohm, I., Antretter, T., Sirén, C., Grichnik, D. & Wincent, J. (2022), 'It's a peoples game, isn't it?! a comparison between the investment returns of business angels and machine learning algorithms', *Entrepreneurship: Theory and Practice* **46**, 1054–1091.

Charbuty, B. & Abdulazeez, A. (2021), 'Classification based on decision tree algorithm for machine learning', *Journal of Applied Science and Technology Trends* **2**(01), 20–28.

Chaudhari, S. L. & Sinha, M. (2021), 'A study on emerging trends in indian startup ecosystem: big data, crowd funding, shared economy', *International Journal of Innovation Science* .

Choudhary, L., Taparia, K., Pandey, A. & Kakkar, A. (2022), 'Indian startup ecosystem 2021'.
URL: *https://hansshodhsudha.com/volume2-issue4/Manuscript%202.pdf*

Corea, F., Bertinetti, G. & Cervellati, E. M. (2021), 'Hacking the venture industry: An early-stage startups investment framework for data-driven investors', *Machine Learning with Applications* **5**, 100062.

Lee, S. H., chul Kim, H. & Lee, K. H. (2018), 'Content-based success prediction of crowdfunding campaigns: A deep learning approach', *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* pp. 193–196.
URL: *https://doi.org/10.1145/3272973.3274053*

Li, J. (2020), Prediction of the success of startup companies based on support vector machine and random forset, *in* '2020 2nd International Workshop on Artificial Intelligence and Education', pp. 5–11.

Lin, C.-Y. (2019), 'Detecting the market reaction of start-ups on gisa equity crowdfunding in taiwan by decision tree algorithm.', *International Journal of Performance Measurement* **9**(2).

Ross, G., Das, S., Sciro, D. & Raza, H. (2021), 'Capitalvx: A machine learning model for startup selection and exit prediction', *The Journal of Finance and Data Science* **7**, 94–114.

Saura, J. R., Reyes-Menéndez, A., Dematos, N. & Correia, M. B. (2021), 'Identifying startups business opportunities from ugc on twitter chatting: An exploratory analysis', *Journal of Theoretical and Applied Electronic Commerce Research* **16**, 1929–1944.

Sharchilev, B., Ozornin, D., Roizner, M., Serdyukov, P., Rumyantsev, A. & Rijke, M. D. (2018), 'Web-based startup success prediction', *International Conference on Information and Knowledge Management, Proceedings* pp. 2283–2292.

Sherk, T., Tran, M. T. & Nguyen, T. V. (2019), 'Sharktank deal prediction: Dataset and computational model', *Proceedings of 2019 11th International Conference on Knowledge and Systems Engineering, KSE 2019* .

Snehal, S., Sundaram, R. & Krishnashree, A. (2020), 'Assessing and comparing top accelerators in brazil, india, and the usa: Through the lens of new ventures' performance', *Entrepreneurial Business and Economics Review* **8**, 153–177.

Taboga, M. (2022), 'Cross-country differences in the size of venture capital financing rounds: a machine learning approach', *Empirical Economics* **62**, 991–1012.

Żbikowski, K. & Antosiuk, P. (2021), 'A machine learning, bias-free approach for predicting business success using crunchbase data', *Information Processing & Management* **58**(4), 102555.

Zhang, Q., Ye, T., Essaidi, M., Agarwal, S., Liu, V. & Loo, B. T. (2017), 'Predicting startup crowdfunding success through longitudinal social engagement analysis'.
URL: *https://doi.org/10.1145/3132847.3132908*

Zhou, J., Gandomi, A. H., Chen, F. & Holzinger, A. (2021), 'Evaluating the quality of machine learning explanations: A survey on methods and metrics', *Electronics* **10**(5), 593.